Aus dem Institut für Medizinische Biometrie und Epidemiologie

der Philipps-Universität Marburg

Direktor: Prof. Dr. rer. nat. Helmut Schäfer

Familienbasierte Assoziationstests bei fehlenden Daten

INAUGURAL-DISSERTATION

zur Erlangung des Doktorgrades der gesamten Medizin dem Fachbereich Humanmedizin der Philipps-Universität Marburg

vorgelegt von

JOCHEN MITTEMEYER

aus Berlin

Marburg 2001

Angenommen vom Fachbereich Humanmedizin
der Philipps-Universität Marburg am 22.11.2001

Gedruckt mit Genehmigung des Fachbereichs

Dekan:	Prof. Dr. Rudolf Arnold
Referent:	PD Dr. Andreas Ziegler
Correferent:	Prof. Dr. Manuela Koch

Inhaltsverzeichnis

1 Einleitung

1.1 Ziel der Arbeit

Seit der Jahrhundertwende werden Krankheiten beschrieben, die in Familien betroffener Personen überproportional häufig auftreten. Daher werden sie vermutlich weitervererbt. Bis heute sind rund 3.000 erbliche Krankheiten bekannt, die auf einen Defekt im Genom beruhen (WHITE & LALOUEL, 1997). Größtenteils sind die Symptome und der Verlauf dieser Krankheiten gut untersucht und beschrieben. Auf welchen der ungefähr 35.000 bis 120.000 humanen Gene (EWING & GREEN, 2000; LIANG *et al.*, 2000) der Defekt für die Krankheit liegt, ist bei einem Großteil der Erbkrankheiten allerdings weiterhin unbekannt (LANDER & SCHORK, 1994). Daher sind Regionen auf den Chromosomen interessant, die bei erkrankten Personen häufiger zu finden sind als bei gesunden Personen.

Das Ziel moderner genetischer Forschung ist, zu diesen Krankheiten entsprechende Varianten der humanen DNA ausfindig zu machen. Mit diesem Wissen erhofft man sich, die Pathogenese der Krankheit zu klären, einen diagnostischen Test zu entwickeln und schließlich eine möglichst kausale Therapie zu finden (WHITE & LALOUEL, 1997). Die Veränderungen in den Genen können Punktmutationen, Deletionen, Insertionen oder Rastermutationen sein. Gendefekte innerhalb von Genabschnitten, die für ein Protein kodieren, können das Protein in seiner Struktur und dadurch in seiner biochemischen Eigenschaft verändern. Mögliche Folgen sind fehlerhafte intrazelluläre oder membranständige Prozesse.

Der Fortschritt der molekularen Genetik ermöglichte, daß seit Anfang der 80er Jahre für viele Erbkrankheiten die Chromosomen lokalisiert wurden, auf denen die defekten Gene liegen. Bis 1995 konnte für mehr als 60 Krankheitsgene die exakte chromosomale Position identifiziert werden (LANDER & KRUGLYAK, 1995). 1989 wurde erstmals der entscheidende Gendefekt der Mukoviszidose entdeckt. Bei der Mukoviszidose (OMIM, 219700) ist der kodierende Bereich auf dem langen Arm des Chromosoms 7 (Genlocus 7q31.2) betroffen. Dieser Bereich kodiert für ein membranständiges Protein, das den transmembranösen Chloridfluß reguliert. In dem Gen, das für dieses Protein mit dem Kürzel CFTR (cystic fibrosis conductance regulator gene) kodiert, fehlen bei vielen Patienten drei Nucleotide. Dadurch sind die Wasser- und Elektrolytströme durch die

Zellmembran gestört, so daß durch sekretorische Dysfunktion ein muköses Sekret entsteht, das vor allem die Bronchialwege verlegt. Wenn nur ein einziger Gendefekt eine Krankheit hervorruft, wird von einer monogenetischen Krankheit gesprochen.

Neben monogenetischen Krankheiten, wie z.B. Hämophilie A/B (OMIM, 306700; OMIM 306900) oder Hämochromatose (OMIM, 235200), gibt es eine größere Anzahl von Krankheiten, wie z.B. Diabetes mellitus Typ II (OMIM, 125853), Krebs oder Herz-Kreislauferkrankungen, die in den wenigsten Fällen auf einen einzelnen Gendefekt zurückgeführt werden können (SIEGENTHALER, 1994). Wirken bei der Manifestation einer Krankheit mehrere Gene mit, spricht man abhängig von der Anzahl der beteiligten Gene von einer oligogenen oder polygenen Krankheit. So werden beispielsweise 18 verschiedene chromosomale Regionen verdächtigt, Gene zu enthalten, die bei der Ätiologie des Diabetes mellitus Typ I eine Rolle spielen (WEEKS & LATHROP, 1995). Spielen neben diesen polygenen Faktoren zusätzlich Umweltfaktoren eine Rolle, handelt es sich um eine multifaktorielle oder komplexe Erkrankung. Die Schwierigkeit, voraussagen zu können, ob bei Vorliegen bestimmter Gene auch wirklich eine Krankheit entsteht, ist vergleichbar mit einer Fußballmannschaft. Selbst wenn ein Trainer die persönlichen Eigenschaften seiner Spieler sehr genau kennt, kann er nicht sicher vorhersagen, ob die Mannschaft als Ganzes im nächsten Spiel gewinnen wird (NEFFE, 1999).

Zur Identifikation von Genloci, die eine Rolle bei der Ausprägung eines komplexen Phänotypen aufweisen, bedient sich die medizinische Genetik verschiedener Methoden. Als Genlocus wird die chromosomale Lage der beiden Kopien eines Gens auf dem entsprechenden Chromosom bezeichnet (SIEGENTHALER, 1994). So ist es möglich, durch Vergleich des Vererbungsmusters einer Krankheit mit den Vererbungsmustern bekannter Genloci, Rückschlüsse auf den unbekannten Krankheitslocus zu ziehen (LANDER & SCHORK, 1994). Mittlerweile sind viele Genloci in ihrer Basensequenz hinreichend analysiert, so daß sie als sogenannte Markerloci verwendet werden können. Markerloci sind Genabschnitte, die mindestens zwei Varianten besitzen. Man spricht von Polymorphismus, wenn mindestens zwei Allele existieren, die mindestens mit einer Frequenz von 1% in der Population vorkommen. In der Humangenetik werden häufig sogenannte Mikrosatelliten verwendet, die hinreichend polymorph sind, so daß die Wahrscheinlichkeit, zufällig heterozygote Personen auszuwählen, hoch ist (STRACHAN & READ, 1999). Bei komplexen Krankheiten wird ein sogenannter Genomscan mit

typischerweise 300-500 genetischen Markern, die über das gesamte Genom verteilt sind, durchgeführt. Indirekt wird versucht, über die bekannte Position der Markerloci den unbekannten Krankheitslocus auf einen kleinen Genabschnitt einzuengen. Durch Feinkartierung wird anschließend die physikalische Region weiter eingegrenzt, so daß danach die genaue Basensequenz des verdächtigen Genabschnitts durch Positionsklonierung bestimmt werden kann. Erstmalig wurde dieses Vorgehen von HORIKAWA *et al.* (2000) erfolgreich bei CAPN10 und Diabetes mellitus Typ II angewendet. Nachdem HANIS *et al.* (1996) durch Kopplungsanalysen unter mexikanischstämmigen US-Amerikanern einen verdächtigen Genlocus, bezeichnete als NIDDM1, im Chromosomenband 2q37.3 gefunden hatten, konnten HORIKAWA *et al.* (2000) durch Feinkartierung den verdächtigen Genabschnitt von 1,7 Mb auf 240 kb eingrenzen. Nach Positionsklonierung wurde das Gen CAPN10 gefunden, das eine Assoziation bei mexikanischstämmigen US-Amerikanern, Nordeuropäern in Finnland mit Diabetes mellitus Typ II zeigte.

Bei der Suche nach Varianten von DNA-Sequenzen, sogenannten Allelen, die bei der Entwicklung einer Krankheit mitwirken, spielen in der genetischen Epidemiologie die Begriffe Assoziation und Kopplung eine große Rolle. Von Assoziation wird gesprochen, wenn in einer Population erkrankte Personen überdurchschnittlich häufig ein spezifisches Allel besitzen. Segregiert dieses Allel zusammen mit der Krankheit in Familien, dann liegt Kopplung vor. Es gibt eine Vielzahl statistischer Verfahren, die auf das Vorliegen von Assoziation und/oder Kopplung testen. In dieser Arbeit werden einige dieser Methoden diskutiert.

Der klassische Ansatz zur Untersuchung einer potentiellen Assoziation ist die Fall-Kontroll Studie. In einer Fall-Kontroll Studie werden die Häufigkeiten eines Markerallels bei erkrankten und nicht erkrankten Personen verglichen. Bei diesem Testverfahren treten jedoch insbesondere bei der Versuchsdurchführung Schwierigkeiten auf, die leicht zu falschen Interpretationen der Ergebnisse führen können (LANDER & SCHORK, 1994). Eines der größten Probleme dabei ist Schichtung in einer inhomogenen Population, die zu einer Scheinassoziation führen kann. Schichtung, auch bezeichnet als Populationsstratifikation, liegt vor, wenn sich in Subpopulationen die relativen Verteilungen gleicher Allele am Markerlocus stark voneinander unterscheiden.

Zur Umgehung des Problems von Scheinassoziation wurden familienbasierte Assoziationsstudien entwickelt. Von näherem Interesse ist in dieser Arbeit der Transmission-Disequilibrium Test (TDT) (SPIELMAN *et al.*, 1993) als ein Test auf Kopplung in Anwesenheit von Assoziation. Dieser Test zeichnet sich durch seine Einfachheit, Robustheit und Eleganz aus (CURTIS, 1997). Seit der ersten Beschreibung dieses Tests hat sich der TDT zu einem weit verbreiteten Kopplungstest avanciert Die Idee des TDT ist, daß bei möglicher Kopplung in Anwesenheit von Assoziation zwischen einem Marker und Krankheitslocus Eltern, die an einem biallelischen Markerlocus heterozygot sind, eines der beiden Allele bevorzugt an ihr erkranktes Kind weitervererbt wird. Bei diesem Ansatz ist keine Kontrollgruppe notwendig, sondern nur ein am Markerlocus vollständig typisiertes Familientrio. Vorteil dieses Tests ist, daß er familienbasiert arbeitet und daher im Gegensatz zu klassischen Assoziationsstudien auch bei Populationsstratifikation ein gültiger Test auf Kopplung ist (EWENS & SPIELMAN, 1995). Probleme treten bei diesem Test jedoch auf, wenn ein Familientrio nicht mehr vollständig typisiert werden kann, da in diesem Fall ein Bias resultiert (CURTIS & SHAM, 1995). Dieses Problem wird um so größer, je später die Manifestation einer Krankheit auftritt, weil z.B. durch den Tod eines Elternteils keine Typisierungen mehr möglich sind.

Zu diesem Problem wurden eine Reihe weiterer Tests entwickelt, die die Idee des TDT verwenden. So benutzt man beim Sib-TDT (SPIELMAN & EWENS, 1998) Geschwister als Kontrollen und untersucht, ähnlich einer gematchten Fall-Kontrollstudie, ob ein erkranktes Geschwisterkind überzufällig häufig ein spezifisches Markerallel besitzt als sein nicht erkranktes Geschwister. Allerdings besitzt dieses Design bei seltenen Krankheitsallelen nur eine geringe Power (ZIEGLER & HEBEBRAND, 1998). Der von KNAPP (1999) vorgestellte RC-TDT verwendet bei fehlenden Daten die Informationen von gesunden Geschwisterkindern, allerdings für einen multiallelischen Marker. Der 1-TDT von SUN *et al.* (1999) ist dagegen gerade bei Elter-Kind Paaren anwendbar, die Situation bei der die Anwendung des TDTs nach CURTIS & SHAM (1995) eigentlich verboten ist. Der Test kann mit dem klassischen TDT kombiniert werden und ist daher ein sehr interessanter und einfacher Ansatz zur Lösung des Problems bei partiell fehlenden Daten.

Ziel dieser Arbeit ist die Untersuchung eines alternativen Ansatzes bei partiell fehlenden Daten. Die zentrale Idee dieses Verfahrens ist die Rekonstruktion, der

fehlenden Genotypen unter Verwendung der vorhandenen Daten. Diese Datenrekonstruktion wird mit dem sogenannten EM-Algorithmus (LAIRD, 1993) durchgeführt. In Schritt 1 wird zuerst eine Schätzung der Allel- und Genotypfrequenzen aus Eltern der vollständigen Familien vorgenommen. In Schritt 2 werden unter Verwendung dieser Frequenzen die Genotypen der fehlenden Eltern auf zwei verschiedene Weisen geschätzt. Bei der ersten Methode werden die beiden fehlenden Allele einzeln geschätzt, und bei der zweiten Methode werden die fehlenden Genotypen direkt geschätzt. Danach erfolgt in Schritt 3 eine Neuschätzung der Allel- und Genotypfrequenzen unter Verwendung aller elterlichen Daten, einschließlich der rekonstruierten Daten. Schritt 2 und 3 werden so oft wiederholt bis die Allel- bzw. Genotypfrequenzen hinreichend stabil geschätzt sind. Bei jedem Elter-Kind Paare wird danach geprüft, werde Genotypen für den fehlenden Elternteil möglich sind. Bei der EM-Allelrekonstruktion werden die Häufigkeiten der möglichen Genotypen unter Verwendung der geschätzten Allelfrequenzen berechnet, bei der EM-Genotyprekonstruktion werden die geschätzten Genotypfrequenzen zugrundegelegt. Der Eintrag in die T/NT-Tabelle erfolgt dann relative zu den entsprechenden Genotypwahrscheinlichkeiten. Dieses Vorgehen ist möglich, da für den Eintrag in die T/NT-Tabelle und für die Berechnung der TDT-Statistik ganze Zahlen nicht zwingend notwendig sind. Die beiden neuen Varianten des EM-Rekonstruktions TDT werden in Monte-Carlo Simulationen untersucht und mit dem 1-TDT verglichen.

Zur Untersuchung der Güte dieses Verfahrens wurde eine Computerprogramm geschrieben, das mit Hilfe von Monte-Carlo Simulationen geeignete Familientrios künstlich generiert, einen beliebigen Anteil an Väter löscht und die fehlenden Daten nach den oben genannten Verfahren rekonstruiert. Zum Vergleich des Ansatzes der EM-Rekonstruktion wurde zusätzlich das Testdesign der sogenannten 1-TDT von SUN *et al.* (1999) in das Programm implementiert. Zur Berücksichtigung des Phänomens von Schichtung werden die Familien aus zwei Subpopulationen mit unterschiedlichen Populationsparametern generiert. Der Anteil der Familien aus den beiden Subpopulationen an der Gesamtpopulation ist dabei variabel. Die Monte-Carlo Simulationen werden sowohl unter den Bedingungen der Nullhypothese bei keiner Kopplung und Assoziation zur Schätzung des empirischen Signifikanzniveaus als auch unter den Bedingungen der Alternativhypothese bei Vorliegen von Assoziation und Kopplung zur Schätzung der Power durchgeführt. Abschließend werden die TDT-

Statistiken für die vollständigen Familien vor dem Löschen der Väter, nach dem Löschen der Väter und nach erfolgreicher Rekonstruktion der fehlenden Daten mit dem EM-Algorithmus berechnet. Die Ergebnisse werden dann zusammen mit den Ergebnissen des 1-TDT untereinander verglichen. Um eine hohe Validität zu erreichen, wurden unter gleichen Bedingungen 10.000 Replikationen durchgeführt.

Zur praktischen Demonstration werden am Ende der Arbeit die beiden EM-Rekonstruktionsmethoden und der 1-TDT auf Datensätze aus zwei Kopplungs- und Assoziationsstudie zwischen Anorexia nervosa sowie Adipositas per magna und zwei Polymorphismen angewendet. Die Datensätze enthalten sowohl Genotypinformation von vollständigen Familientrios als auch von Paaren aus einem erkranktem Kind und einem Elternteil (HINNEY *et al.*, 1997a; HINNEY *et al.*, 1997b).

1.2 Aufbau der Arbeit

In Kapitel 2, 3 und 4 werden die Grundbegriffe der genetischen Epidemiologie, die Mendelschen Regeln, Kopplung und Assoziation, erläutert. Zusätzlich werden in Kapitel 3 klassische Kopplungsstudien vorgestellt.

In Kapitel 5 wird anhand der klassischen Fall-Kontroll Studie das Odds Ratio und der χ^2-Test erläutert, die einen Hinweis auf Assoziation geben. Außerdem werden Gründe erläutert, die zu Scheinassoziation führen können, insbesondere durch Schichtung in einer Population. Nachfolgend werden Empfehlungen für die Durchführung von Assoziationsstudien gegeben.

Kapitel 6 beschreibt sogenannte familienbasierte Assoziationstest, das Haplotype Relative Risk (HRR) und den Transmission-Disequilibrium Test (TDT), die nicht mehr sensibel auf Schichtung sind. Eine zentrale Rolle nimmt dabei der TDT ein, da er auch auf Kopplung in Anwesenheit von Assoziation testet. Weiter wird erläutert, daß ein Bias entsteht, wenn für den TDT Familien mit teilweise fehlenden elterlichen Informationen verwendet werden. Danach werden sieben Testmethoden, der Sib-TDT, SDT, der RC-TDT, die PRG, der LRT und der 1-TDT, die jeweils unter bestimmten Bedingungen verwendet werden können, wenn Daten fehlen.

In Kapitel 7 wird eine neue Idee zur Rekonstruktion der fehlenden Daten, die sogenannte EM-Rekonstruktion, vorgestellt. Für die EM-Rekonstruktion werden zwei Möglichkeiten vorgeschlagen. Die erste Möglichkeit rekonstruiert den fehlenden Genotypen auf der Basis der Allel, die zweite Möglichkeit auf der Basis der Genotypen.

Zur näheren Untersuchung der EM-Rekonstruktion durch Monte-Carlo Simulationen wird ein Computerprogramm beschrieben, das geeignete Familientrios generiert, partiell Daten löscht und mit der EM-Rekonstruktion rekonstruiert. Außerdem wird der 1-TDT in das Computerprogramm implementiert.

In Kapitel 8 werden die Ergebnisse der Monte-Carlo Simulationen sowohl unter der Nullhypothese bei fehlender Assoziation oder Kopplung als auch unter der Alternativhypothese bei Assoziation und Kopplung dargestellt. In den Monte-Carlo Simulationen wird zusätzliche Schichtung in der Gesamtpopulation berücksichtigt. Außerdem wird die Anwendung der EM-Rekonstruktion und des 1-TDT anhand von Originaldaten demonstriert.

Die Ergebnisse der MC-Simulationen werden in Kapitel 9 diskutiert. Dazu werden die Ergebnisse der EM-Rekonstruktion mit den Ergebnissen des 1-TDT in bezug auf das empirische Signifikanzniveau und die Power verglichen. Am Ende wird ein Ausblick auf weitere interessante Fragestellungen zum Problem von fehlenden Daten bei familienbasierten Assoziationstests gegeben.

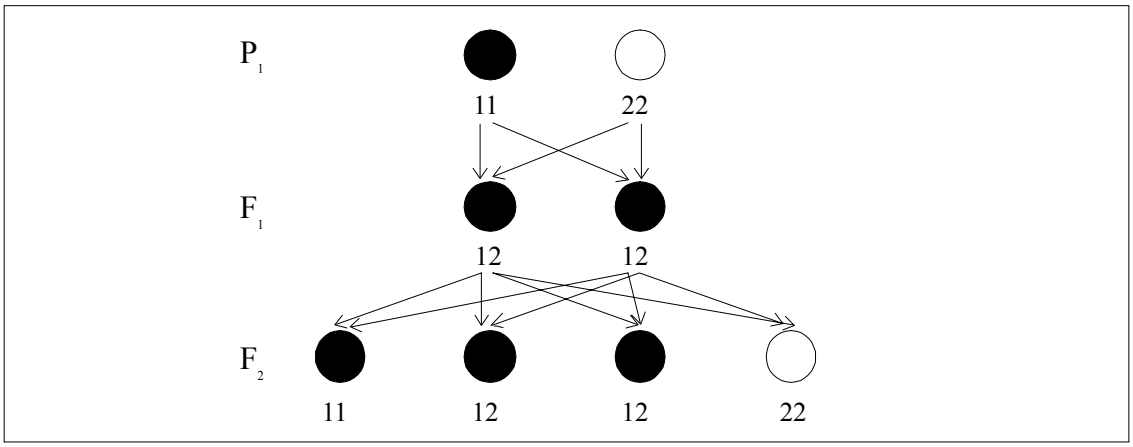
2 Mendelsche Vererbung

Die folgende Darstellung der klassischen Genetik lehnt sich an die Lehrbücher *Genetik* von STRICKBERGER (1988) und *Human Molecular Genetics* von STRACHAN & READ (1999) an. In der Genetik ist die Beobachtung und Vererbung biologischer Merkmalen an die nächste Generation von großer Bedeutung. Georg Mendel (1822-1884) beobachtete als erster systematisch die Vererbung von Merkmalen bei der Erbse Pisum sativum. Mendel wählte sieben Eigenschaften für seine Beobachtungen aus (Samenform, Samenfarbe, Hülsenfarbe, etc.). Bei seinen Kreuzungsversuchen machte Mendel folgende Beobachtungen und stellte danach die nach ihm benannten Regeln auf.

2.1 Spaltungsregel und Uniformitätsregel

Mendel kreuzte Erbsensorten mit gelbem und grünem Samen und erhielt nur Erbsen mit gelbem Samen. Die Erbsen der paternalen Ausgangskreuzung werden P_1-Generation, ihre Nachkommenschaft F_1-Generation und alle nachfolgenden Generationen F_2-Generation usw. benannt. Als er dann die Erbsen der F_1-Generation mit sich selbst bestäubte, erhielt er wieder Erbsen mit beiden Merkmalen der P_1-Generation. Das Verhältnis der Merkmale in der F_2-Generation lag sehr nahe bei 3:1.

Abbildung 2-1: Erbschema für ein Beobachtungsmerkmal
Farbe: schwarz kodiert durch Allel 1 und weiß durch Allel 2
Allel 1 ist dominant über Allel 2



Ein Merkmal, das in der F_1-Generation verschwindet, tritt in der nachfolgenden Generation in einem Viertel der Fälle wieder auf. Daraus leitet sich ab, daß Merkmale zwar verborgen bleiben können, jedoch nicht zerstört werden. Ein Merkmal, das ein anderes Merkmal unterdrückt, wird als voll dominant bezeichnet. Ein Merkmal, das in dessen Anwesenheit unterdrückt wird, wird als voll rezessiv bezeichnet. Später wurde

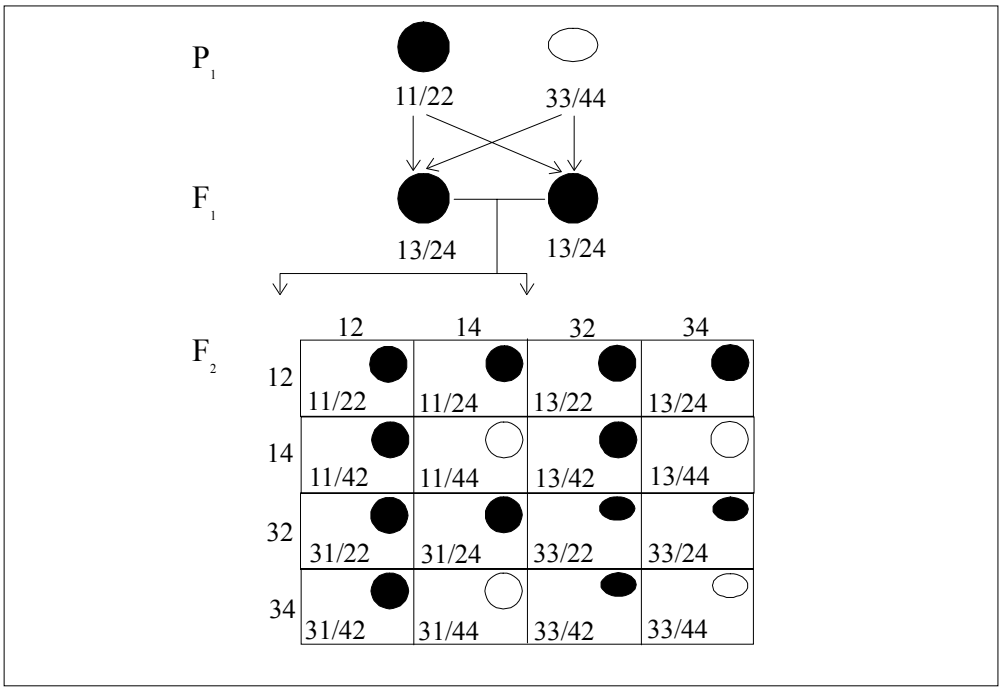
entdeckt, daß der für ein Merkmal kodierende Genlocus auf der DNA-Doppelhelix zweimal angelegt ist. Die beiden Varianten eines Genlocus werden Allele genannt (SIEGENTHALER, 1994). Besitzt ein Individuum am Genlocus zwei identische Allele, wird es homozygot genannt. Sind zwei verschiedene Allele anwesend, wird es als heterozygot bezeichnet. Mendel stellte weiterhin keine Unterschiede fest, ob die Merkmale von der väterlichen oder der mütterlichen Seite vererbt wurden. Eine solche Vererbung wird autosomal genannt.

Aus diesem Versuch läßt sich als weiteres die Uniformitätsregel ableiten. Wenn bei einem beliebigen Kreuzungsversuch alle Individuen der F_1-Generation identisch für das beobachtete Merkmal sind, so kann angenommen werden, daß die Elterngeneration am entsprechenden Genlocus homozygot ist.

2.2 Unabhängigkeitsregel

Nach Entdeckung der Gesetzmäßigkeiten über die Aufspaltung zwischen zwei Varianten eines Merkmals, untersuchte Mendel weiterhin die Vererbung bei Individuen, die sich in zwei Merkmalen unterschieden. Dazu kreuzte Mendel Erbsen mit glatten und gelben Samen mit Erbsen, die runzelige und grüne Samen bildete. Nach dem folgenden Kreuzungsschema erhielt er Erbsen vom Typ glatt/gelb, glatt/grün, runzelig/gelb und runzelig/grün im Verhältnis 9:3:3:1.

Abbildung 2-2: Erbschema für zwei Beobachtungsmerkmale
Form: rund kodiert durch Allel 1 oder oval kodiert durch Allel 3
Farbe: schwarz kodierte durch Allel 2 und weiß kodiert durch Allel 4



(Alle 1 ist dominant über Allel 3 und Allel 2 ist dominant über Allel 4)

Die Verhältnisse der einzelnen Merkmalsarten, glatt oder runzelig sowie gelb oder grün, betragen auch in diesem Fall jeweils 3:1. Kombiniert man die Vererbung beider Merkmalsarten, stellt man fest, daß sich beide Merkmalstypen unabhängig voneinander verhalten. Um die Wahrscheinlichkeit für das Auftreten einer bestimmten Merkmalskombination zu erhalten, müssen die jeweiligen Einzelwahrscheinlichkeiten multipliziert werden.

Bei den oben beschriebenen einfachen genetischen Merkmalen entscheidet allein der Genotyp am Genlocus, welches Merkmal zur Ausprägung kommt. Für einige Stoffwechselerkrankungen gibt es einen einzigen Genotyp, der allein für das Auftreten des Merkmals bzw. der Krankheit verantwortlich ist. Ein solches Merkmal wird Mendelsches Merkmal genannt. Bei menschlichen Merkmalen sind jedoch meistens mehrere Genloci, individuelle Faktoren und Umwelteinflüsse bei der Ausprägung beteiligt. Ist eine Krankheit von vielen Faktoren abhängig, folgt die Vererbung der einzelnen DNA-Sequenzvarianten zwar formal den Mendelschen Regeln, die Krankheit selbst jedoch nicht. Solche Krankheiten werden auch komplexe Krankheiten genannt. Komplexität kann durch Umweltfaktoren oder durch den unterschiedlichen Einfluß eines oder mehrerer Genloci bedingt sein (STRACHAN & READ, 1999).

2.3 Abweichungen von den Mendelschen Regeln

Wenn ein Genlocus nicht wie bisher angenommen auf einem autosomalen Chromosom, sondern auf einem Gonosom liegt, dann ist der Genlocus bei dem männlichen Karyotyp XY nur einmal vorhanden. Allele auf diesem Genlocus müssen daher nicht mit einem zweiten Allel konkurrieren und sind folglich immer dominant. Häufigstes Beispiel eines solchen X-chromosomalen Erbgangs ist die Hämophilie A und B. Von dieser Krankheit sind fast ausschließlich männliche Patienten betroffen, da bereits ein Kopie des defekten X-Chromosoms zur Erkrankung führt. Weibliche Personen mit dem Karyotyp XX erkranken bei einer Kopie des defekten X-Chromosoms nicht, da der Defekt rezessiv ist. Heterozygote Frauen werden als Konduktoren bezeichnet.

Eine weitere Ursache für Abweichungen von den Mendelschen Regeln kann Phänokopie sein. Bestimmte Umwelteinflüsse können einen Organismus derart verändern, daß der Phänotyp die Wirkung eines Gens simuliert. Ein solches Individuum wird als Phänokopie bezeichnet. Ein Beispiel ist der insulinabhängige Diabetes mellitus Typ I. Die Entfernung, Verletzung oder Entzündung der Bauchspeicheldrüse kann bei genotypisch normalen Personen die identische Klinik eines klassischen Diabetes mellitus Typ I hervorrufen. Auch andere Umwelteinflüsse wie Medikamente, Virusinfektionen, Chemikalien oder Radioaktivität können Phänokopien erzeugen.

3 Kopplung

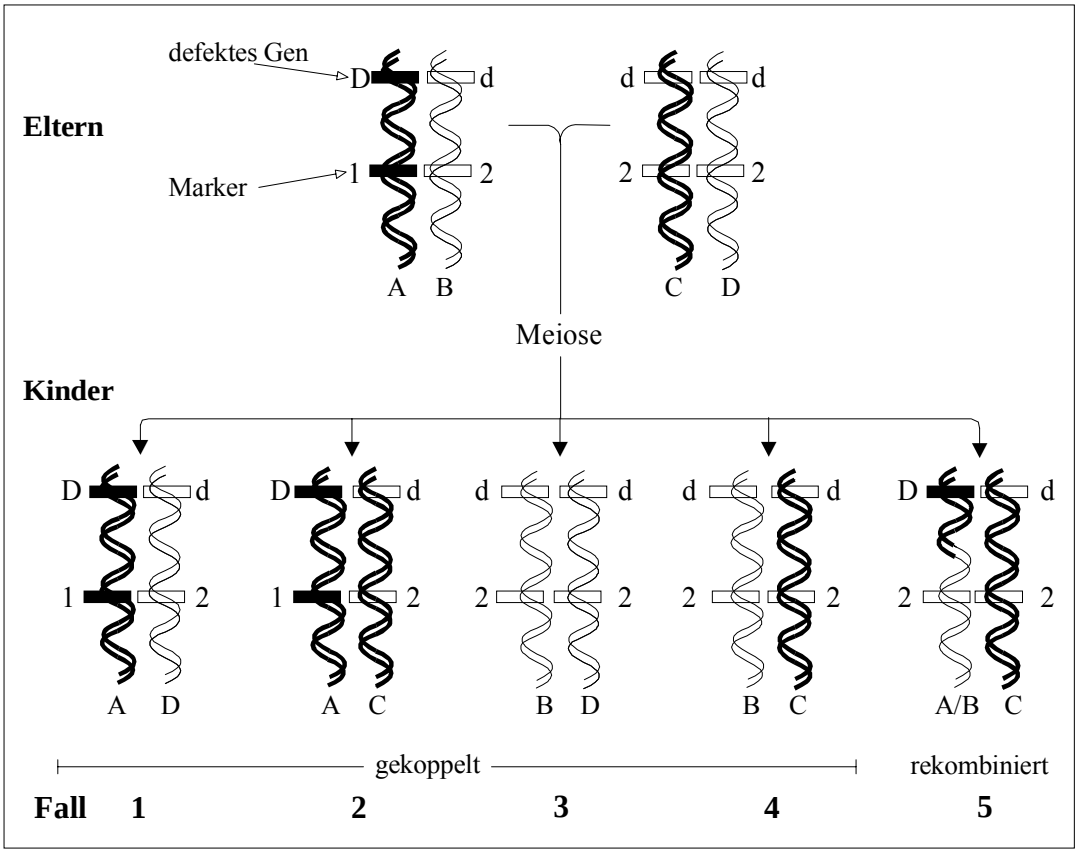
Bei der Suche nach chromosomalen Regionen, die eine funktionsrelevante DNA-Variante enthalten, werden in der Genetischen Epidemiologie zwei verschiedene Ansätze verwendet. Der eine Ansatz beruht auf Kopplung, der andere Ansatz auf Assoziation. Beide messen auf unterschiedliche Weise eine Abweichung von den Mendelschen Regeln der unabhängigen Vererbung. In diesem Kapitel wird zuerst der Begriff der Kopplung erklärt, und es werden in der Genetischen Epidemiologie häufig verwendete Studien vorgestellt, die auf Kopplung testen. Das nächste Kapitel beschäftigt sich mit dem Begriff der Assoziation.

3.1 Rekombination

Nach den Mendelschen Regeln sind Gene unabhängig und frei kombinierbar. Die Regel der freien Kombinierbarkeit der Gene ist jedoch nicht unbeschränkt gültig. Die moderne Genetik betrachtet Gene nicht mehr isoliert, sondern berücksichtigt auch deren chromosomale Lage. Wenn zwei Genloci auf zwei verschiedenen Chromosomen liegen, werden beide mit einer Wahrscheinlichkeit von 50% zusammen vererbt. Wenn zwei Genloci allerdings gemeinsam auf einem Chromosom liegen, sind die Genloci theoretisch nicht mehr unabhängig. Vermutlich müßten sie immer gemeinsam gekoppelt vererbt werden. Diese Annahme mißachtet jedoch ein Phänomen das Crossing-over genannt wird. Bei der Vererbung kommt es in der Prophase der Meiose I zu meiotischer Rekombination von DNA-Abschnitten. Das heißt, homologe Chromosomenpaare lagern sich aneinander, und es kommt zu einem Austausch von DNA-Abschnitten (STRACHAN & READ, 1999). Kopplung zwischen Genloci auf einem Chromosom ist also nicht dauerhaft und kann durch Crossing-over unterbrochen werden. Thomas H. Morgan (1866-1945) entdeckte dieses Phänomen bei Kreuzungsversuchen mit der Fruchtfliege Drosophila.

In der nachfolgenden Abbildung sind verschiedene Möglichkeiten bei der Vererbung von zwei benachbarten Genloci dargestellt. In den Fällen 1 bis 4 werden bei der Meiose die Chromosomen A bis D als Ganzes vererbt, d.h. es liegt eine gekoppelte Vererbung vor. Im Fall 5 dagegen findet zwischen den Chromosomen A und B eine Rekombination statt, die zu dem Austausch der Marker 1 und 2 führt.

Abbildung 3-1 Gekoppelte und ungekoppelte Vererbung



(angelehnt an Bild 3, WHITE & LALOUEL, 1997)

Ein Rekombination kann dann erkannt werden, wenn zwischen zwei Genloci ein einziges Crossing-over stattfindet. Finden dagegen zwischen zwei Genloci zwei Crossing-overs statt, wird zwar ein Stück des DNA-Strangs zwischen den beiden Genloci ausgetauscht, aber dieser Vorgang läßt sich durch alleinige Betrachtung der Genloci nicht von der gekoppelten Vererbung unterscheiden. Rekombinationen bleiben also unsichtbar, wenn eine gerade Anzahl von Crossing-overs stattfinden. Die Wahrscheinlichkeit für mehrere Rekombinationen zwischen zwei Genloci wird zwar immer geringer, aber es stellt eine Fehlerquelle dar, weil Rekombinationsraten dadurch unterschätzt werden können. Unter Rekombination wird daher ein sichtbarer Austausch homologer Chromosomenabschnitte während der Meiose verstanden. Diese Situation gilt für eine ungerade Anzahl an Crossing-overs. Eine Rekombination zwischen zwei Genloci wird um so seltener, je enger beide auf einem Chromosomenabschnitt nebeneinander liegen, da der Abschnitt in dem die Rekombination stattfinden muß, um beide Loci zu trennen, sehr klein ist. Ein Block von eng benachbarten Allelen wird als Haplotype bezeichnet (STRACHAN & READ, 1999, S. 270).

Wenn Haplotypen nicht durch Crossing-over ausgebrochen werden, werden sie oft gemeinsam innerhalb von Stammbäumen vererbt. Bei Kopplung zwischen einem Markerallel und einem Krankheitsallel wird davon ausgegangen, daß beide auf der DNA so nah beieinander liegen, daß sie „nur selten“ nicht gemeinsam vererbt werden. Die Rekombinationsrate θ ist die Wahrscheinlichkeit, daß zwischen beiden Loci eine ungerade Anzahl von Crossing-overs stattfindet. Je kleiner also die Rekombinationsrate ist, um so näher sollten die beiden Allel auf dem Chromosomenstrang beieinanderliegen.

3.2 Genetische Marker

Um innerhalb von Stammbäumen bzw. Familien Allele erkennen zu können, die gemeinsam mit einer Krankheit vererbt (co-segregiert) werden, sollte die Mehrzahl der Personen heterozygot am Markerlocus sein. Nur bei heterozygoten Personen kann eine Rekombination beobachtet werden, denn eine homozygote Person vererbt immer dasselbe Markerallel. Deshalb sind hinreichend polymorphe genetische Marker notwendig, für die eine zufällige ausgewählte Person mit einer hohen Wahrscheinlichkeit heterozygot ist. Außerdem sollte die Typisierung billig und schnell aus leicht zu gewinnendem Material, z.B. Blut, erfolgen können. Die Entwicklung von Markersystem verlief über Blutgruppen und Polymorphismen von Serumproteinen zu den heute verwendeten DNA-VNTR Mikrosatelliten und single nucleotide polymorphisms (SNPs).

3.3 Kopplungsstudien

In Kopplungsstudien wird getestet, ob innerhalb einer Familie ein Markerallel überzufällig häufiger zusammen mit einer Krankheit vererbt wird. Bei der Konstruktion eines Kopplungstests wird bei sogenannten modellbasierten Methoden (z.B. die klassische LOD-Score Methode) ein bestimmtes Vererbungsmodell oder bei modellfreien Methoden kein Vererbungsmodell berücksichtigt werden (ELSTON, 1998).

Gerade bei sogenannten komplexen Krankheiten wird eine modellbasierte Kopplungsanalyse durch unvollständige Penetranzen, genetische Heterogenität, variables Manifestationsalter, Phänokopie oder diagnostische Ungenauigkeiten schnell sehr schwierig (NÖTHEN *et al.*, 1992). Alle Faktoren bei der Auswahl des oft großen Studienkollektivs zu berücksichtigen, ist häufig aussichtslos. Bei den modellbasierten Methoden besteht außerdem die Gefahr, bei der Annahme eines falschen

Vererbungsmodells ein falsch positives Ergebnis zu erhalten, so daß dann das Testergebnis nicht interpretierbar wird (ELSTON, 1998). Aufgrund dieser Schwierigkeiten haben sich die modellfreien Methoden etabliert. Der große Vorteil der modellfreien Methoden ist, daß keine Spezifikation des zugrundeliegenden genetischen Modells erforderlich ist. Der prominenteste Vertreter der modellfreien Methoden ist die affected sib pair (ASP) Methode. Dabei werden Familien mit zwei oder mehreren erkrankten Geschwistern rekrutiert. Gegenstand der Kopplungsstudie ist die Frage, ob diese Paare ein bestimmtes Markerallel nach Herkunft, d.h. identical by descent (IBD), gar nicht, einfach oder doppelt gemeinsam tragen (ELSTON, 1998). Die in den ASP beobachtete Verteilung der IBD-Werte wird dann unter Verwendung eines statistischen Tests mit derjenigen verglichen, die unter der Nullhypothese „keine Kopplung“ erwartet wird. Bekannte Teststatistiken hierfür sind der sogenannte MLS-Test (RISCH, 1990) und der Mean-Test (BLACKWELDER & ELSTON, 1985), die beide Optimalitätseigenschaften besitzen.

4 Assoziation

Der zweite Ansatz bei der Suche nach funktionsrelevanten DNA-Varianten basiert auf dem Begriff von Assoziation. Kopplung und Assoziation sind grundsätzlich zwei unterschiedliche Phänomene. Bei Kopplung wird eine Abweichung von den Mendelschen Regeln innerhalb von Stammbäumen gemessen. Dagegen findet man bei Assoziation im Gruppenvergleich von Kranken und Gesunden ein spezifisches Allel häufiger bei Kranken. Für Assoziation gibt es viele mögliche Gründe, die nicht immer genetisch begründet werden können.

4.1 Beziehung zwischen Allelen

Angenommen in einer Population sei p_1 der Anteil von Personen mit Diabetes mellitus Typ I und p_2 der Anteil von Personen mit dem HLA-DR3-Antigen (RIEDE & SCHÄFER, 1999). Der Anteil der Personen mit Diabetes mellitus Typ I und dem HLA-DR3-Antigen wäre erwartungsgemäß bei Unabhängigkeit der beiden Merkmale das Produkt aus p_1 und p_2 (OTT, 1991). Allerdings kommt aus zunächst unbekanntem Grund das HLA-DR3-Antigen überzufällig häufig zusammen mit der Diabetes mellitus Typ I vor.

Angenommen die Wahrscheinlichkeit, daß das HLA-DR3-Antigen auftritt, sei P(1), und die Wahrscheinlichkeit für das Auftreten des Diabetes mellitus Typ I sei P(D). Für das gleichzeitige Auftreten von 1 und D wird die Wahrscheinlichkeit P(D1) angenommen. Wenn in einer Population beide Merkmale voneinander abhängig sind, weicht das Produkt der Einzelwahrscheinlichkeiten von der Wahrscheinlichkeit des Auftretens beider Genloci um die Differenz δ ab.

$$P(D1) = P(D) \cdot P(1) + \delta$$

$$\delta = P(D1) - P(D) \cdot P(1)$$

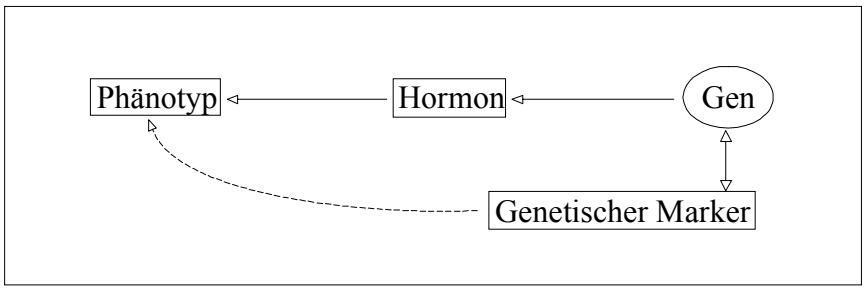
Bei ungleichmäßiger Verteilung wird von Assoziation zwischen beiden Merkmalen gesprochen. Der Wert des Assoziationsparameters δ ist ein Maß für das Verteilungsungleichgewicht zwischen den beiden Merkmalen. δ gibt sowohl Größe als auch die Richtung der Assoziation an.

Nimmt δ den Wert 0 an, liegt keine Assoziation vor. Nimmt δ einen Wert größer Null an, liegt eine positive Assoziation vor. Beide Merkmale treten häufiger zusammen auf, als unter Unabhängigkeit zu erwarten wäre. Das HLA-DR3-Antigen stellt damit einen

Risikofaktor dar. Das heißt, bei Auftreten des HLA-DR3-Antigen ist die Chance, an Diabetes mellitus Typ I zu erkranken, erhöht. Nimmt δ einen Wert kleiner Null an, liegt eine negative Assoziation vor. Die Chance, sowohl ein bestimmtes Allel als auch die Krankheit zu besitzen, ist kleiner als die Chance unter Unabhängigkeit.

Da in den dargestellten Beispielen das Krankheitsgen nicht bekannt ist und beobachtet werden kann, ist der einzige Hinweis auf die Anwesenheit des vermeintlichen Krankheitsgens der Phänotyp der Krankheit. Bei einer genetischen Krankheit wird davon ausgegangen, daß ein defektes Gen z.B. für ein defektes Hormon kodiert und folglich den Phänotyp einer Krankheit verursacht. Auf der Suche nach dem unbekannten Krankheitsgen ist es sinnvoll, nach genetischen Merkmalen zu suchen, die eine Assoziation mit dem Phänotyp aufweisen. Wird eine Assoziation beobachtet, kann auch vermutet werden, daß auch eine Beziehung zwischen dem genetischen Marker und dem Krankheitsgen besteht.

Abbildung 4-1 Schema der Ursache-Wirkungs-Beziehung



Der wahre Grund für Assoziation zwischen Markerlocus und dem Phänotyp ist jedoch weiterhin ungewiß. Interessant für die genetische Forschung sind Assoziationen, die auf Kopplung beruhen. Kopplung liegt vor, wenn der Markerlocus in dichter Nachbarschaft oder innerhalb des Krankheitsgen liegt. Wird der genetische Marker verdächtig, selber das Krankheitsallel zu sein, wird von einem Kandidatengen gesprochen. Es gibt jedoch auch Gründe, daß eine Assoziation besteht, ohne daß der Markerlocus in der Nähe des Krankheitsgen liegt.

4.2 Faktoren für Assoziation

Wie bereits erwähnt, besagt eine positive Assoziation nur, daß zwei Merkmale innerhalb einer Population häufiger zusammen auftreten, als statistisch zu erwarten wäre. Es gibt für positive Assoziation Faktoren, die für die Erforschung einer Krankheit brauchbar sind, aber auch Faktoren, die zu einer sogenannten Scheinassoziation führen.

Daher ist das Wissen über die möglichen Ursachen von Assoziation sowohl bei der Interpretation von Assoziationsstudien als auch bei der Entwicklung neuer statistischer Verfahren wichtig.

4.2.1 Enge Kopplung

Das Allel am Markerlocus liegt in enger Nachbarschaft oder innerhalb des Krankheitsgens und die genetische Kopplung bleibt über mehrere Generationen von Rekombinationen erhalten (OWEN *et al.*, 1997). Dieser Grund ist interessant für Assoziationsstudien, weil enge Kopplung einen Hinweis auf die Position des Krankheitsgens gibt.

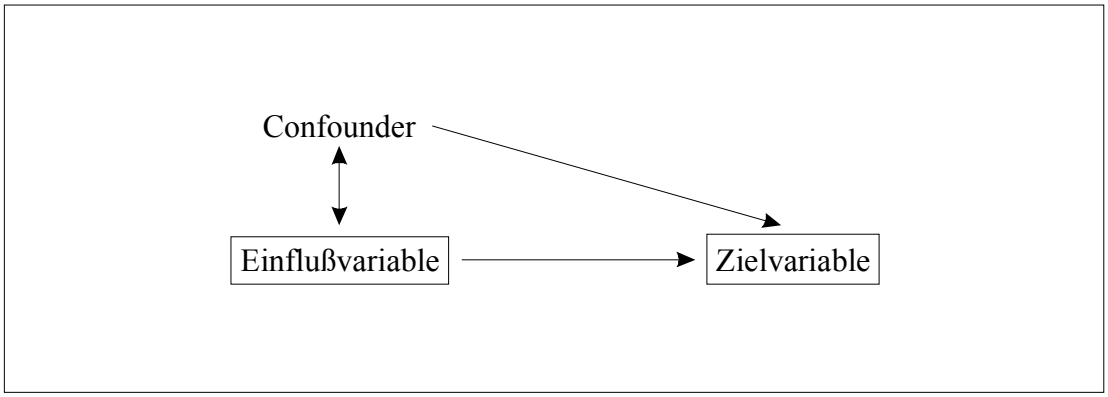
4.2.2 Markerlocus identisch mit Krankheitsgen

Es ist auch möglich, daß der Markerlocus direkt dem relevanten Krankheitsallel entspricht. Eine derartige Assoziation müßte dann in allen Populationen gefunden werden, in denen diese Mutation auftritt (LANDER & SCHORK, 1994). Diese Möglichkeit ist der Idealfall für Assoziationsstudien, da eine Assoziation nicht durch Rekombination aufgehoben werden kann.

4.2.3 Confounder

Es gibt jedoch auch andere Faktoren für Assoziation, die sich nicht auf Kopplung zurückführen lassen. Grund für falsch positiven Assoziationen sind Störvariablen, sogenannte Confounder. Ein Confounder wirkt kausal auf die Zielvariable, ist nicht Ziel der Untersuchung, aber assoziert mit der Zielvariable Krankheit. Typische Confounder sind Alter, Rauchen und Geschlecht.

Abbildung 4-2 Interaktion zwischen Confounder und der Zielvariablen



(nach KREIENBROCK & SCHACH, 2000)

Ein weiterer wichtigster Grund für falsch positive Ergebnisse ist jedoch ein Phänomen, das auf Populationsstratifikation beruht. Populationsstraktifikation ist identisch zu Confounding durch Ethnizität. Ein Beispiel für Populationsstratifikation aus der Literatur (LANDER & SCHORK, 1994) ist die Assoziation zwischen der Fähigkeit, mit Stäbchen essen zu können, und dem Vorkommen des HLA-A_1-Antigens. Wenn in San Francisco ohne Berücksichtigung der ethnischen Gruppen Fälle und Kontrollen rekrutiert würden, wäre bei dieser Studie eine positive Assoziation zwischen Stäbchenessen und dem HLA-A_1-Antigen zu erwarten. Der Fehler der Studie bestände allerdings darin, daß die untersuchte Population in sich nicht homogen ist. Weil San Francisco eine multikulturelle Stadt mit einem hohen Anteil von Personen asiatischer Abstammung ist, würden sowohl europäische und afrikanische Amerikaner, als auch asiatische Amerikaner in die Studie einbezogen. Da in der asiatischen Subpopulation kulturell bedingt mit Stäbchen gegessen wird und das Markerallel HLA-A_1 ethnologisch in dieser Gruppe erhöht ist, würde in der Studie eine Assoziation gefunden, die auf Schichtung zurückzuführen ist. In diesem Fall wäre die ethnische Herkunft (Asiat ja/nein) der Confounder, das HLA-A_1-Allel die Einflußvariable und das Stäbchenessen die Zielvariable. Populationsstratifikation ist also Confounding durch ethnische Heterogenität. Assoziationsstudien sollten deshalb innerhalb relativ homogener Populationen durchgeführt werden.

Es gibt ein einfaches Beispiel, anhand dessen anschaulich Schichtung erklärt werden kann. Man stelle sich zwei Populationen mit einem Anteil von jeweils 50% an der Gesamtpopulation und den Frequenzen P(A) und P(a) für die Allele A und a sowie P(B) und P(b) für die Allele B und b vor. Unter Hardy-Weinberg Bedingungen seien die

beiden Genloci frei kombinierbar. Daraus ergeben sich folgende Wahrscheinlichkeiten für die Genotypen an beiden Genloci.

Tabelle 4-1 Beispiel für Schichtung bei zwei Populationen mit einem Anteil von 50% an der Gesamtpopulation und unterschiedlichen Allelfrequenzen für die Allel A und a sowie die Allele B und b.

Population 1	**Population 2**
P(A)=P(a)=0,5	P(A)=1
P(B)=1	P(B)=P(b)=0,5
$\Rightarrow$ P(AB)=P(aB)=0,5	$\Rightarrow$ P(AB)=P(Ab)=0,5
P(Ab)=P(ab)=0	P(aB)=P(ab)=0

Werden nun beide Populationen in einem Verhältnis von 50:50 gemischt werden, ergeben sich folgende Wahrscheinlichkeiten für die Genotypen an den beiden Genloci.

P(AB)=0,5 P(Ab)=P(aB)=0,25 P(ab)=0

In der Gesamtpopulation haben die beiden Allele A und B folgende Frequenzen.

P(A)=0,75 P(B)=0,75

Nach der Definition von Assoziation ergibt sich daraus:

$$\delta = P(AB)-P(A)\cdot P(B)$$

$$= 0,25-0,75\cdot 0,75$$

$$= -0,0625$$

Obwohl in den beiden Populationen 1 und 2 Hardy-Weinberg Bedingungen gelten, d.h. keine Assoziation vorliegt, ergibt sich in der Mischpopulation eine scheinbare Abweichung von der freien Kombinierbarkeit der beiden Genloci. In diesem Fall beruht die positive Assoziation auf Schichtung.

4.2.4 Multiples Testen

Bei Suchtests mit vielen verschiedenen Markerloci, die über das gesamte Genom verteilt sind, erhöht sich die Wahrscheinlichkeit für einen Fehler 1. Art (LANDER & KRUGLYAK, 1995). Bei einem Suchtest mit 20 verschiedenen Markern ist bei einem nominellen Signifikanzniveau von 5% im Schnitt mit einem falsch positiven Ergebnis zu rechnen (OWEN *et al.*, 1997). Deshalb ist in diesem Fall eine Bonferroni-Korrektur

(CHRISTENSEN, 1996) oder eine Bonferroni-Holm-Korrektur (SACHS, 1974) notwendig. Die Notwendigkeit von Korrekturen wird von PATERSON (1997) dargestellt. Bei 5 vermuteten Krankheitsgenen für eine neurologische Krankheit und 20.000 potentiellen Kandidatengenen sind bei einem geforderten p-Wert von 0,05 bis zu 99,5% der als signifikant deklarierten Assoziationen falsch positiv. Bei einem geforderten p-Wert von 0,001 liegt die Rate falsch positiver Assoziationen immer noch bei 80%. Erst für p-Werte $<10^{-8}$ könnten verwertbare Aussagen getroffen werden.

Um dieses Problem zu umgehen, wird von verschiedenen Autoren ein Sample-splitting empfohlen (OWEN *et al.,* 1997). Dazu wird die Studiengruppe in zwei Untergruppen geteilt, und ein Assoziationstest mit einem nicht zu strengen nominellen Signifikanzniveau in der ersten Gruppe durchgeführt. Positive Ergebnisse werden anschließend in der zweiten Gruppe mit einem korrekten nominellen Signifikanzniveau überprüft.

5 Fall-Kontroll Assoziationsstudien

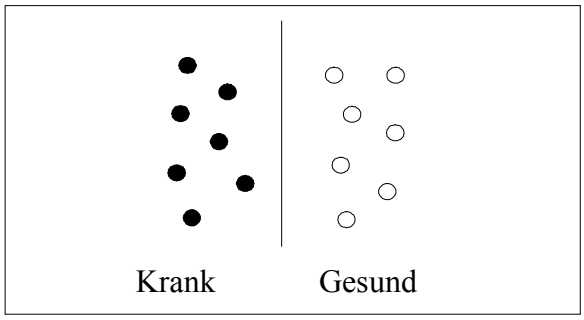
In Assoziationsstudien wird untersucht, ob ein bestimmtes Allele eines bekannten Markerlocus (Kandidatengen) gehäuft mit einer bestimmten Erkrankung auftreten. In der Annahme, daß ein assoziierter Markerlocus in der Nähe oder innerhalb des vermutlichen Krankheitsgen liegt, kann sich die weitere Suche auf einen kleineren Genomabschnitt beschränken (OWEN *et al.*, 1997). Sogenannten Kandidatengene werden bevorzugt Genregionen entnommen, die einen biochemischen Bezug zur Krankheit haben. Bei Assoziationsstudien zum Diabetes mellitus Typ I ist beispielsweise die Genregion interessant, die für das Insulinprotein codiert.

Es gibt eine Reihe von Methoden, mit denen auf Assoziation getestet werden kann. Einige gebräuchliche Verfahren werden im folgenden dargestellt. Bei sogenannten Beobachtungsstudien wird die Beziehung zwischen einer Krankheit und einem oder mehreren Risikofaktoren untersucht. Bei der Kohortenstudie wird eine Population über einen bestimmten Zeitraum hinsichtlich des Auftretens einer Erkrankung prospektiv beobachtet. Aufgrund eines meist langen Beobachtungszeitraumes ist eine derartige Studie sehr zeit- und kostenintensiv. Bei der Fall-Kontroll Studie werden kranke Personen und gesunde Personen werden hinsichtlich ihrer Exposition zu einem früheren Zeitpunkt miteinander verglichen. Der Auswahl der Kontrollgruppe kommt dabei große Bedeutung zu (KREIENBROCK & SCHACH, 2000). Die Kontrollgruppe sollte bei genetischen Studien aus einer homogenen Population stammen und zufällig ausgewählt sein. Das ist notwendig, weil stark voneinander abweichende Genfrequenzen am gleichen Genlocus, wie bereits in Kapitel 4.2 erläutert, in einer gemischten Population zu falsch positiven Ergebnissen führen können.

5.1 Odds Ratio

Zur Erläuterung wird eine Fall-Kontroll Studie mit folgender Ausgangssituation gewählt. Die eine Gruppe umfaßt kranke Personen (K=1) und die andere Gruppe gesunde Personen (K=0).

Abbildung 5-1 Aufbau einer Fall-Kontroll Studie



In beiden Gruppen wird die Zahl der Personen bestimmt, die mit einem Risikofaktor exponiert sind (E=1) oder nicht exponiert sind (E=0). Das Ergebnis wird in Form einer Vierfeldertafel, der sogenannten Kontingenztabelle, dargestellt.

Tabelle 5-1 Beobachtete Anzahlen von Erkrankten und Gesunden unter exponierten und nicht exponierten Personen

	Exponiert	**Nicht exponiert**
Krank	n_1	n_2
Gesund	n_3	n_4

Eine Möglichkeit, einen Vergleich zwischen zwei Gruppen bezüglich der Erkrankung herzustellen, basiert auf dem sogenannten Odds. Das Odds ergibt sich im Gegensatz zur Wahrscheinlichkeit aus dem Begriff der Chance. Allgemein definiert man als Odds einer Wahrscheinlichkeit P den Ausdruck

$$\text{Odds}(P) = \frac{P}{1 - P}.$$

Aus diesem Ausdruck ergibt sich die Chance, mit der ein Ereignis eintritt. Bei einer Wahrscheinlichkeit von P=0,5 ergibt sich ein Odds(P)=0,5/0,5=1 oder eine Chance von 1:1. Bei P=0,75 erhält man ein Odds(P)=0,75/0,25=3 oder eine Chance von 3:1 für das Eintreten des Ereignisses.

Bei einem Vergleich von zwei Chancen kann untersucht werden, ob eine Exposition einen Effekt auf die Entstehung einer Krankheit hat. Das Odds Ratio (OR) ist definiert als der Quotient aus der Chance, daß exponierte Personen erkranken, und der Chance, daß nicht exponierte Personen erkranken.

$$OR = \frac{Odds(P(K=1 \mid E=1))}{Odds(P(K=1 \mid E=0))} = \frac{\frac{P(K=1 \mid E=1)}{P(K=0 \mid E=1)}}{\frac{P(K=1 \mid E=0)}{P(K=0 \mid E=0)}} = \frac{n_1 \cdot n_4}{n_2 \cdot n_3}$$

Intuitiv scheint das Odds Ratio wenig effektiv, hat aber in der Epidemiologie eine überaus große Bedeutung. Das Odds Ratio wird als Faktor interpretiert, um den die Chance bei Exposition zu erkranken steigt. Allerdings kann das Odds Ratio auch unter der Frage angewandt werden, wie die Expositionschance steigt, wenn man erkrankt ist. Diese Feststellung ist wichtig, denn damit ist für das Odds Ratio keine Longitudinalstudie zwingend notwendig. Das Odds Ratio kann deshalb auch in Fall-Kontroll Studien angewandt werden, da es auch als Verhältnis der Expositionschance interpretiert werden kann. In Tabelle 5-2 wird ein fiktives Bespiel für Einträge in die Kontingenztabelle gegeben.

Tabelle 5-2 Beispiel für Einträge in die Kontingenztabelle bei einer Fall-Kontroll Studie

	Exponiert	**Nicht exponiert**	Summe
Krank	30	10	40
Gesund	20	40	60
Summe	50	50	100

Die Berechnung des Schätzers für das Odds Ratio erfolgt wie bereits beschrieben.

$$\hat{OR} = \frac{30 \cdot 40}{10 \cdot 20} = 6$$

In diesem Fall steigt die Chance bei kranken Personen exponiert zu sein um den Faktor 6 gegenüber gesunden Personen.

Das Relative Risiko (RR) ist dagegen als als Quotient der Risiken unter Exposition und keiner Exposition definiert (WOOLF, 1955). Das Relative Risiko ist damit ein multiplikativer Faktor, um den sich das Risiko zu erkranken erhöht, wenn eine Person exponiert ist. Ist das RR=1, so sind beide Wahrscheinlichkeiten identisch. Ist das Relative Risiko RR>1, so liegt die Wahrscheinlichkeit zu erkranken unter exponierten Personen höher als unter nicht exponierten Personen.

Unter gewissen Umständen kann davon ausgegangen werden, daß das Odds Ratio und das Relative Risiko nahezu identisch sind. Für seltene Erkrankungen, Faustregel <10%,

bei denen das Odds sowohl unter Exponierten als auch unter nicht Exponierten sehr klein ist, sind Odds Ratio und Relatives Risiko ungefähr gleich. Es ist bei dieser Annahme jedoch stets zu prüfen, ob die betrachtete Krankheit wirklich selten ist.

Die bisher erläuterten Maßzahlen beschreiben für eine betrachtete Gesamtheit von Testpersonen jedoch nur Unterschiede zwischen zwei Gruppen. Die epidemiologische Frage nach einer möglichen Ursache-Wirkungs-Beziehung wird dadurch nicht beantwortet. Denn auch nach der Durchführung der Studie bleibt die wahre Größe der erhobenen Maßzahl und der Grund für das Ergebnis unbekannt.

5.2 χ^2-Test

Die Idee, daß eine Exposition eine Erkrankung tatsächlich beeinflußt, kann durch statistische Tests untermauert werden. In der statistischen Nomenklatur wird von der Alternative H_1, es liegt ein Einfluß vor, und der Nullhypothese H_0, es liegt kein Einfluß vor, gesprochen. Hierbei können zwei mögliche Fehler auftreten. Ein Fehler 1. Art tritt ein, wenn die Alternative angenommen wird, obwohl sie in Wahrheit nicht vorliegt. Wird die Alternative abgelehnt, obwohl sie in Wahrheit gilt, wird von einem Fehler 2. Art gesprochen. Abhängig vom Studienumfang gibt es zwei wesentliche Verfahrenstypen zur Beantwortung der Fragestellung. Bei kleinen Studien müssen exakte Testverfahren angewandt werden, während bei hinreichend großen Studien ein asymptotischer χ^2-Test möglich ist.

Der χ^2-Test ist anwendbar für die bereits beschriebene Kontingenztabelle (vgl. Tabelle 5-2). Innerhalb dieser Tabelle testet der χ^2-Test auf Homogenität. Der χ^2-Test wird anhand des Beispiels aus Tabelle 5-2 beschrieben.

Bei Unabhängigkeit der Krankheit von der Exposition wäre eine Verteilung zu erwarten, bei der die Zahlen der Exponierten und nicht Exponierten in den Gruppen der Kranken und Gesunden gleich sind.

Tabelle 5-3 Erwartete Einträge in die Kontingenztabelle bei Unabhängigkeit

	Exponiert	**Nicht exponiert**	Summe
Krank	20	20	40
Gesund	30	30	60
Summe	50	50	100

Es kann nun getestet werden, ob statistisch signifikante Abweichungen von der unabhängigen Verteilung vorhanden sind. Dazu wird folgende Teststatistik verwendet, die ein Schätzer für die Abweichung von der Unabhängigkeitsverteilung ist.

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Dabei gibt O_i die Anzahl der beobachteten Einträge in den einzelnen Zellen aus Tabelle 5-2 an und E_i die entsprechend zu erwartenden Einträge in den einzelnen Zellen bei Unabhängigkeit aus Tabelle 5-3.

Als Faustregel gilt, daß in jeder Zelle der Unabhängigkeitstabelle mindestens 5 Einträge erwartet werden müssen, damit das Ergebnis bei einem Signifikanzniveau von minimal 1% mit einer χ^2-Verteilung mit 1 Freiheitsgrad verglichen werden kann. Für das Beispiel aus Tabelle 5-2 ist eine möglich Teststatistik:

$$T = \frac{(30-20)^2}{20} + \frac{(10-20)^2}{20} + \frac{(20-30)^2}{30} + \frac{(40-30)^2}{30} = 16.7$$

Liegt der Testwert über der 95%-Quantile von 3,8415, wird das Ergebnis als statistisch signifikant gewertet. Aus einer Verteilungstabelle für die χ^2-Verteilung kann für den Testwert T der entsprechende empirische p-Wert abgelesen werden. In diesem Fall p = 0,0000439.

Da das geforderte Signifikanzniveau von 5% unterboten wurde, kann mit einer Wahrscheinlichkeit von 5% für einen Fehler 1. Art angenommen werden, daß bei Exposition die Erkrankung signifikant häufiger auftritt. Ergebnisse aus Fall-Kontroll Studien sind jedoch selten eindeutig, da verschiedene Untersucher bei der gleichen Fragestellung zu widersprüchlichen Ergebnissen kommen können. Wie bereits in Kapitel 4.2 beschrieben, gibt es eine Reihe von Faktoren, die zu unterschiedliche Ergebnisse führen können.

5.3 Empfehlungen für Assoziationsstudien

Assoziationsstudien bergen sowohl für den Untersucher als auch für eine wissenschaftliche Fachzeitschrift, in dem eine positive Assoziation veröffentlicht wird, die Gefahr eines späteren Vertrauensverlusts, wenn im nachhinein die Assoziation nicht bestätigt werden kann oder sogar widerlegt wird. Daher werden unter anderem von wichtigen wissenschaftlichen Fachzeitschriften an den Untersucher und die Studie

hohen Anforderungen bei der Planung und Durchführung gestellt. Eine Minimierung des Risikos einer Scheinassoziation durch Populationsstratifikation kann durch eine sorgfältige Auswahl der Kontrollen erreicht werden (OWEN *et al.*, 1997; BARON, 1997). Dennoch bleibt auch bei größter Sorgfalt das Risiko für ein falsch positives Ergebnis. Daher sollte die Fallzahlen groß sein, da insbesondere bei komplexen Krankheiten der Geneffekt oft klein ist. Außerdem sollte der p-Wert klein sein, das Markerallel in einem plausiblen biologischen Verhältnis zur untersuchten Krankheit stehen und eine geeignete statistische Analyse durchgeführt werden. Zusätzlich wird bei eine gefundenen Assoziation eine Bestätigung in einer zweiten unabhängigen Studie gefordert, wobei Assoziation einmal in einer familienbasierten Assoziationsstudien beobachtet werden sollte (EDITORIAL *Nature Genetics*, Mai 1999). Schließlich sollten negative Studien in verschiedenen ethischen Populationen kritisch beurteilt werden, da eine Assoziation auch nur in einer ethischen Population vorkommen kann (NIMGAONKAR, 1997).

5.4 Genom-Kontroll-Tests

Der Vorteil von Fall-Kontroll Studien ist, daß sie einfach und billig durchzuführen sind und sind deshalb eine attraktive Methode, um Assoziation zu finden. Dennoch besteht die Gefahr, bei Schichtung in der Population ein falsch positive Ergebnisse zu erhalten. Aus diesen Überlegungen heraus wurde von PRITCHARD & ROSENBERG (1999) eine neue Idee vorgestellt. Dabei wird sowohl in einer Fall-Kontroll Studie auf Assoziation zwischen einem Kandidatengen und dem Phänotyp einer Krankheit getestet, als auch im nächsten Schritt bei positivem Ergebnis in der Population nach Schichtung gesucht. Zu diesem Zweck werden Fälle und Kontrollen möglichst ohne offensichtliche Schichtung rekrutiert. Dann werden für eine geringe Anzahl von Markerloci die Allelfrequenzen ermittelt. Ist ein Markerlocus signifikant mit dem Phänotyp assoziiert, wird mit einer Gruppe von ungekoppelten Marker auf Populationsstratifikation getestet. Bei der Verwendung von Mikrosatelliten wird eine notwendige Anzahl von ≥15-20 angegeben, wenn ein nominales Signifikanzniveau von 5% gefordert wird. Bei der Verwendung von 30 biallelischen Marker kann das empirische Signifikanzniveau bis zu 6% erreichen.

Im Zuge des Human Genom Project ist bisher die erste Generation von biallelischen Markern, sogenannte single nucleotide polymorphisms (SNPs), entdeckt worden, die sich nur in einem Nucleotid unterscheiden, hinreichend polymorph sind und in dichten Abständen über das ganze Genom verteilt sind (COLLINS *et al.*, 1998; WANG *et al.*,

1998). Anfang 2001 wurde bereits eine Karte von 1,42 Millionen SNP mit durchschnittlich einem SNP je 1,9 Kilobasen beschrieben (SACHIDANANDAM *et al.*, 2001). Diese SNP-Karte ermöglicht in Zukunft weiterentwickelte Studiendesigns und neuen Analyseverfahren mit denen Assoziation auch bei komplexen Erkrankungen schnell und sicher entdeckt kann (Böddeker & ZIEGLER, 2000; RISCH & MERIKANGAS, 1996). Da sich biallelische SNP-Marker nur in einem einzigen Nucleotid unterscheiden, sind sie einer automatisierten Analyse über Chips leicht und schnell zugänglich. Allerdings können sich bei der Analyse Probleme durch multiples Testen und mögliche Fehltypisierungen ergeben.

Eine Assoziationsstudie, die bereits diesen Ansatz verfolgt, stammt von DEVLIN & ROEDER (1999). Wenn Populationsstratifikation vorliegt, so überschätzt eine Teststatistik Assoziation zwischen einem Markerallel und einer Krankheit. Unter einem vernünftigen genetischen Modell in der Population kann jedoch angenommen werden, daß die Überschätzung von Assoziation über das gesamte Genom nahezu konstant bleibt. Diese Beobachtung ermöglichte Bayes´sche Korrektur für die Fall-Kontroll Studie. Je größer die Überschätzung von Assoziation ist, desto größer wird auch der Verlust an Power nach der Korrektur. Daher sollten auch hier die Fälle und Kontrollen sorgfältig ausgesucht werden.

6 Familienbasierte Assoziationsstudien

Assoziationsstudien stellen ein große Herausforderung an den Untersucher insbesondere bei der Auswahl der Kontrollgruppe. Bei einer schlechten Auswahl der Kontrollpersonen, besonders in ethnische inhomogen Populationen, läuft der Untersucher Gefahr eine Assoziation zu finden, die in Wahrheit auf Schichtung in der Population zuzuführen ist. Daher ist es erstrebenswert eine Testmethode zu haben, bei der die Kontrollpersonen aus der gleichen ethnischen Population stammen, wie die erkrankten Personen. In diesem Kapitel werden zwei familienbasierte Assoziationstests, das sogenannte Haplotype Relative Risk (HRR) (FALK & RUBINSTEIN, 1987) und der Transmission-Disequilibrium Test (TDT) (SPIELMAN *et al.*, 1993) erläutert, die zur Vermeidung von Populationsstratifikation ein entsprechendes Matching zwischen Fällen und Kontrollen dadurch erreicht, daß die Häufigkeiten, mit ein bestimmtes Allel vererbt bzw. nicht vererbt wird, miteinander verglichen werden.

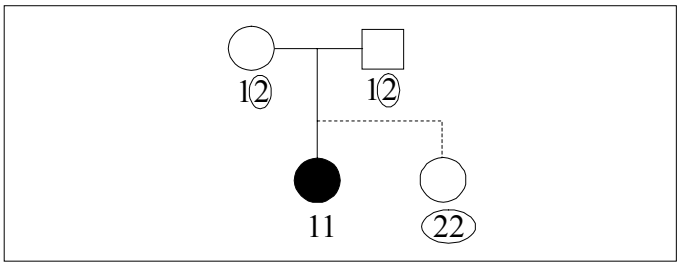
6.1 Interne Fall-Kontroll Studie

Aus den bereits genannten Gründen, die zu Scheinassoziation führen können, ist ein Verfahren sinnvoll, das auf eine reale Kontrollgruppe verzichtet. Dennoch sollte das Verfahren in der Lage ist, eine Aussage zu treffen, ob bei Exposition das Erkrankungsrisiko steigt. Eine mögliche Lösung des Problems geben FALK & RUBINSTEIN (1987) mit dem Haplotype Relative Risk (HRR). Die HRR Methode ist eine Alternative zu dem Relativen Risiko von WOOLF (1955), das Risiko einer Erkrankung in Abhängigkeit von der Anwesenheit eines bestimmten Markerallels zu schätzen (OTT, 1989).

Das HRR verzichtet auf eine reale Kontrollgruppe und „rekrutiert“ als Kontrollen die beiden nicht vererbten Allele der Eltern. Die Summe der nicht vererbten Allelpaare bildet eine fiktive Kontrollgruppe (LANDER & SCHORK, 1994). Vorteil diese Vorgehens ist, daß die möglichen Probleme durch eine inadäquate Auswahl der Kontrollpersonen vermieden werden. Denn die fiktive Kontrollgruppe ist exakt aus der selben Gruppe entnommen, wie die Fallgruppe. Dadurch wird Confounding bei ethnischer Heterogenität vermieden. Außerdem erfolgt die Typisierung immer im selben Labor. Außerdem können Fehltypisierungen teilweise durch Vergleich der Segregationsmuster bei Mendelscher Vererbung erkannt werden.

Nicht mehr anwendbar ist das HRR Verfahren allerdings bei Krankheiten, die sich erst im fortgeschrittenen Alter manifestieren. Das gilt besonders für degenerative Erkrankungen wie Morbus Alzheimer oder Morbus Parkinson. Bei diesen Krankheiten ist es teilweise schwer oder sogar unmöglich, komplette Familientrios zu bekommen, weil die Eltern verstorben sind. Außerdem ist ein Mehraufwand zur Typisierung der Testpersonen um 50% notwendig, da beim HRR für ein Fall-Kontroll Paar drei anstatt zwei Personen notwendig sind.

Abbildung 6-1 Bildung einer internen Kontrollgruppe, Allel 1 wird zweimal vererbt, Allel 2 wird zweimal nicht vererbt und bildet die fiktive Kontrolle



Aus einer Familie erhält man sowohl Informationen für die Fallgruppe als auch für die interne Kontrollgruppe. In einer Vierfeldertafel wird dann für jede Familie eingetragen, ob das erkrankte Kind und die interne Kontrollgruppe das Markerallel 1 oder Allel 2 besitzen. Pro Vererbung ergeben sich zwei Einträge in die Vierfeldertafel. Bei einer Anzahl von n Familientrios und vier Einträgen pro Familie werden also 4n Einträge vorgenommen.

Tabelle 6-1 Vierfeldertafel mit der Anzahl der Fälle, in denen das erkrankte Kind und die interne Kontrolle das Allel 1 oder Allel 2 besitzen

	Allel 1	**Allel 2**	Anzahl
Fall	w	x	2n
Interne Kontrolle	y	z	2n

Das HRR vergleicht die Frequenz der Anwesenheit von Allel 1 in der Fallgruppe mit der Frequenz des Allels 1 in der internen Kontrollgruppe. Damit entspricht die Form des HRR der Berechnung des RR und des OR (KNAPP *et al.*, 1993). Ähnlich dem RR ist das HRR bei keiner Assoziation nicht von 1 verschieden.

$$HRR = \frac{w/x}{y/z} = \frac{wz}{xy}$$

Allerdings hat das HRR die Eigenschaft das RR immer zu unterschätzen und damit immer dichter an 1 zu liegen als das RR (KNAPP *et al.*, 1993).

$$|HRR-1| \leq |RR-1|$$

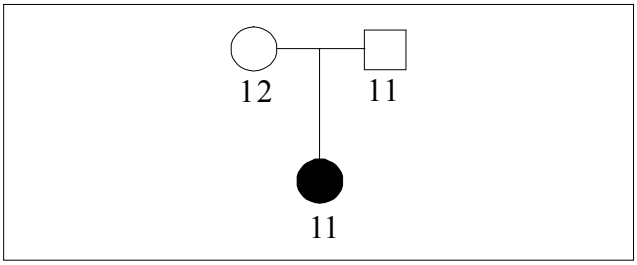
OTT (1989) benutzt in seiner Arbeit eine andere Form der Darstellung, die eine gepaarte Information erfaßt. Jedes Elternteil transmittiert (T) ein Allel an das Kind, und ein Allel wird nicht transmittiert (NT). Werden diese beiden Beobachtungen zusammen erfaßt, so spricht man von gepaarter Information. Zu diesem Zweck hat OTT (1989) die sogenannte T/NT-Tabelle eingeführt (vgl. Tabelle 6-2). Im Gegensatz zur Arbeit von FALK & RUBINSTEIN (1987) wird nun keine Kontrollgruppe mehr gebildet. Der Eintrag für das transmittierte und das nicht transmittierte Allel erfolgt in einem Feld.

Tabelle 6-2 T/NT-Tabelle nach OTT (1989)

Transmittiert	**Nicht transmittiert**	
	1	2
1	a	b
2	c	d

An einem binären Markerlocus ergeben sich vier mögliche Kombinationen. Ist ein Elternteil am Markerlocus homozygot, ist das transmittierte Allel gleich dem nicht transmittierten Allel. Der entsprechender Eintrag erfolgt im Feld a oder d. Ist ein Elternteil am Markerlocus heterozygot, wird entweder das Allel 1 oder das Allel 2 transmittiert, und das jeweils andere Allel nicht transmittiert. Wird das Allel 1 transmittiert und das Allel 2 nicht transmittiert, erfolgt der Eintrag im Feld b. Ist die Situation umgekehrt, erfolgt der Eintrag im Feld c.

Abbildung 6-2 Beispiel für eine fiktive Familie mit einem erkrankten Kind



Die entsprechenden Einträge für die fiktive Familien werden in Tabelle 6-3 demonstriert. Bei der ersten Transmission wird Allel 1 transmittiert und Allel 2 nicht

transmittiert, der Eintrag erfolgt in Zelle b. Bei der zweiten Transmission wird Allel 1 transmittiert und Allel 1 nicht transmittiert, der Eintrag erfolgt in Zelle a.

Tabelle 6-3 Einträge in die T/NT-Tabelle für die fiktive Familie aus Abbildung 6-2

Transmittiert	**Nicht transmittiert**	
	1	2
1	|	|
2		

OTT (1989) hat weiterhin in einem rezessiven Modell die theoretisch zu erwartenden relativen Häufigkeiten für die vier möglichen Kombinationen berechnet. Die Frequenz des Markerallels 1 sei m, die Frequenz des Krankheitsallels D sei p, der Assoziationsparameter sei δ, und die Rekombinationsrate sei θ. Die relativen Häufigkeiten gelten für ein Elternteil.

Tabelle 6-4 Relative Häufigkeiten der Vererbung elterlicher Allele am Markerlocus an ein erkranktes Kind bei einem rezessiven Vererbungsmodell in Abhängigkeit von den Frequenz p des Krankheitsallels D, der Frequenz m des Allels 1, der Rekombinationsrate θ und des Assoziationsparameters δ für ein Elternteil.

Transmittiert	**Nicht transmittiert**		**Summe**
	1	2	
1	$(m+\delta/p)m$	$(m+\delta/p)(1-m)-\theta\delta/p$	$m+(1-\theta)\delta/p$
2	$(1-m-\delta/p)m+\theta\delta/p$	$(1-m-\delta/p)(1-m)$	$1-m-(1-\theta)\delta/p$
Summe	$m+\theta\delta/p$	$1-m-\theta\delta/p$	1

Berechnet man mit diesen relativen Häufigkeiten das HRR, so wird das HRR unter der folgende Bedingung gleich 1.

$$\delta(1-2\theta) = 0$$

$$\delta = 0 \text{ oder } \theta = ½.$$

Anders ausgedrückt ist das HRR nur dann von 1 verschieden, wenn $\delta \neq 0$ und $\theta \neq ½$ ist. Das HRR ist damit unter einem rezessiven Modell nicht wie das RR sensibel für Scheinassoziation und überschätzt Assoziation nicht. Das RR wird dagegen schon von 1 verschieden, wenn unabhängig von eventuell vorliegender Kopplung oder Schichtung in der Population der Assoziationsparameter $\delta \neq 0$ ist (OTT, 1989). Allerdings dürfen nur

Familien verwendet werden, die nicht aus Stammbäumen stammen, die über mehrere Generationen von der Krankheit betroffen sind. Das HRR ist nur ein Test auf Assoziation, wenn die Familien unabhängig sind. In einer Population, in der Schichtung vorliegt, bestehen diese Bedingungen erst wieder nach zwei Generationen (EWENS & SPIELMAN, 1995). Sind diese Bedingungen jedoch erfüllt, hat das HRR eine sehr große Power, Assoziation nachzuweisen (EWENS & SPIELMAN, 1995). KNAPP *et al.* (1993) zeigt weiterhin, daß diese Eigenschaft des HRR auch für alle anderen Vererbungsmuster gilt. Dennoch ist das HRR kein direkter Test auf Kopplung SPIELMAN *et al.* (1993). Weil Assoziation aus den genannten Faktoren nur ein indirekter Beweis für Kopplung ist, sind zusätzliche Tests auf Kopplung notwendig (LANDER & SCHORK, 1994).

Nachteil des HRR gegenüber dem klassischen Fall-Kontroll Ansatz ist eine geringere Power, Assoziationen in einer Population ausfindig zu machen (KNAPP *et al.*, 1993). Dieser Nachteil wird jedoch aufgrund seiner Vorteile in der Praxis gerne in Kauf genommen. Einen weiteren Nachteil dieser Methode zeigt folgende Überlegung. Ist ein Elternteil unter einem rezessiven Modell homozygot am Markerlocus, transmittiert es in jedem Fall den Markerlocus an das Kind und an die Kontrollgruppe. Ein derartige Transmission ist nicht informativ, weil ein homozygotes Elter für die Fallgruppe und Kontrollgruppe das gleiche Ergebnis beiträgt. Dennoch trägt diese Transmission zum Testwert bei. Das HRR überschätzt zwar nicht den Wert der Assoziation (KNAPP *et al.*, 1993), aber mit steigender Anzahl von homozygoten Elternteilen wird eine mögliche Kopplung verdeckt (LANDER & SCHORK, 1994). Außerdem unterscheidet der Test nicht, ob ein Elternteil erkrankt ist oder nicht. Wenn beide Eltern und das Kind heterozygot sind, ist die Vererbung ebenfalls nicht informativ, weil jeweils jedes Markerallel einmal transmittiert und einmal nicht transmittiert wird. Es kommt zwar zu keiner Verzerrung, weil die Einträge auf den Nebendiagonalen ausgeglichen sind, aber die Power wird verringert (STRACHAN & READ, 1999).

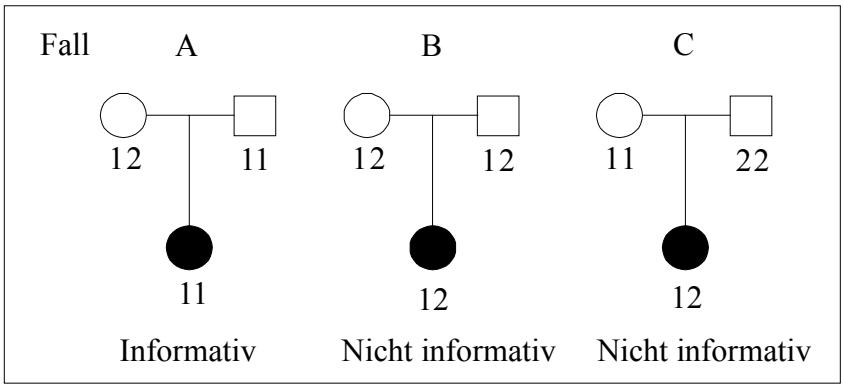
In der gleichen Arbeit von FALK & RUBINSTEIN (1987) wurde mit dem HRR eine positive Assoziation zwischen den DR3- und DR4-Allelen am HLA-DR-Locus und dem Diabetes mellitus Typ I (IDDM) nachgewiesen. Für das DR3-Allel wurde ein Testwert HRR = 4,23 berechnet, das entspricht einem p-Wert von $2{,}6 \cdot 10^{-5}$. Für das DR4-Allel ergab sich ein Testwert HRR = 9,21 und ein p-Wert von $7{,}6 \cdot 10^{-10}$.

6.2 Transmission-Disequilibrium Test

Nachteil bei der Verwendung des HRR ist die Tatsache, daß das HRR nicht die Information benutzt, daß die vererbten und nicht vererbten Allele gepaart vorkommen. Zur Erfassung der gepaarten Information hat bereits OTT (1989) die T/NT-Tabelle vorgestellt (vgl. Tabelle 6-2). Die statistischen Möglichkeiten dieses Ansatzes für einen Test auf Kopplung wurden jedoch erst von SPIELMAN *et al.* (1993) in Form des Transmission-Disequilibrium Tests (TDT) genutzt. Dieser Test nimmt eine gewisse Zwischenstellung ein. Ursprünglich wurde er als ein Test auf Kopplung in Anwesenheit von beliebiger Assoziation entwickelt (SPIELMAN *et al.*, 1993). Andere Autoren sehen den TDT allerdings eher als einen Test auf Assoziation in Anwesenheit von Kopplung (ELSTON, 1998). Trotz der Uneinigkeit bezüglich der statistischen Bewertung des TDT, hat sich dieser Test in den letzten Jahren zu einem etablierten Test auf Kopplung entwickelt, stark beachtet, intensiv weiterentwickelt und häufig in der Praxis eingesetzt wird.

Der TDT erfordert ein Triodesign aus einem erkrankten Kind und beiden Eltern. Alle Familienmitglieder müssen am Markerlocus vollständig typisiert sein. Grundgedanke des TDT ist, daß ein Markerallel, das mit einem Krankheitsallel gekoppelt ist, überzufällig häufig gemeinsam mit dem Krankheitsallel an ein erkranktes Kind vererbt wird. Denn bei Kopplung zwischen einem Markerallel und einem Krankheitsallel würde das Markerallel häufiger als mit einer 50:50 Chance transmittiert. Ist das spezifisches Markerallel zusätzlich auch noch mit dem Krankheitsallel assoziiert, sollte es auch überzufällig häufig bei allen erkrankten Kindern vorkommen. Wenn keine Kopplung vorliegt, wird das Markerallel wie homologe Chromosomen nach den Mendelschen Regeln mit einer Wahrscheinlichkeit von ½ weitervererbt.

Abbildung 6-3: Informative und nicht informative Vererbung bei Familientrio mit einem erkrankten Kind



Informativ sind für den TDT allerdings nur heterozygote Eltern am Markerlocus, denn nur dann kann ein Unterschied bei der Zellteilung beobachtet werden (vgl. Abbildung 6-3, Fall A). Sind jedoch beide Eltern und das Kind heterozygot, ist die Vererbung formal uninformativ, da nicht entschieden werden kann welches Allel vom Vater und von der Mutter vererbt wird (vgl. Abbildung 6-3, Fall B). Dieser Fall kann dennoch für den TDT verwendet werden, da es für den TDT unerheblich ist, ob ein Allel vom Vater oder der Mutter vererbt wird. Eltern, die für den Marker homozygot sind, können dagegen keine Informationen für den Test geben, weil sie unabhängig von Kopplung und Assoziation immer das gleiche Allel an ihr Kind vererben und daher keine Kopplungsinformationen tragen (vgl. Abbildung 6-3, Fall C).

Der TDT setzt dort an, wo auf Populationsebene eine Assoziation zwischen einem Markerallel und einer Krankheit gefunden wurde (SPIELMAN *et al.*, 1993). Der TDT nutzt zum einen die Assoziationsinformation der elterlichen Allele, die nicht an das erkrankte Kind vererbt werden. Zum anderen benutzt er zusätzlich die Kopplungsinformation bei der Verteilung der elterlichen Allele während der Vererbung (Segregation). Der TDT bestätigt also nur dann ein positives Ergebnis von Assoziationsstudien, wenn die vorher gefundene positive Assoziation nicht auf Schichtung oder auf anderen falsch positiven Faktoren beruht. Sind die Genotypen des Familientrios am Markerlocus vollständig bestimmt, so werden sie in die von OTT (1989) vorgestellte T/NT-Tabelle eingetragen (vgl. Tabelle 6-2).

Aus dieser Tabelle wird die TDT-Statistik berechnet. Unter den Bedingungen von Unabhängigkeit sind die Testwerte entsprechend einer χ^2-Verteilung mit einem Freiheitsgrad verteilt.

$$\chi^2 = \frac{(b-c)^2}{b+c}$$

Bei genauer Betrachtung der TDT Statistik lassen sich folgende Beobachtungen machen.

1. Die Teststatistik wird gleich Null, wenn die Einträge in den Felder b und c auf der Nebendiagonalen der T/NT-Tabelle gleich sind (vgl. Tabelle 6-2).
2. Bildet man die Differenz aus b und c erhält man das

   $$\text{Ergebnis}\,(b-c) = 2\frac{\delta}{p}(1-2\theta)$$

3. Die Gleichung in 2. wird gleich Null unter der Bedingung

   $(b\text{-}c)=0 \Leftrightarrow \delta=0 \vee \theta=½$

Die erste Bedingungen für die Nullhypothese H_0 ist erfüllt, wenn keine Assoziation vorliegt $\delta=0$. Die zweite Bedingung ist erfüllt, wenn keine Kopplung nachgewiesen werden kann $\theta = ½$. Ist also mindestens ein Bedingung erfüllt, entweder $\theta = ½$ oder $\delta=0$, wird die TDT-Statistik immer gleich null. Scheinassoziation reicht in Abwesenheit von Kopplung daher nicht aus, daß die TDT-Statistik von der χ^2-Verteilung abweicht. Die alternative Hypothese H_1 wird erst angenommen, wenn zwischen Krankheitsgen und Markerlocus Kopplung und Assoziation besteht. Weicht das Testergebnis statistisch signifikant von der χ^2-Verteilung ab, wird die alternative Hypothese angenommen. EWENS & SPIELMAN (1995) haben für verschiedene Vererbungsmodelle theoretisch gezeigt, daß der TDT ein gültiger Test auf Kopplung in Anwesenheit von Assoziation ist, egal auf welchem Grund die Assoziation beruht. Entsprechend ist der TDT ein zulässiger Kopplungstest in potentiell geschichteten Populationen, in denen nicht Hardy-Weinberg Bedingungen gelten. Wenn allerdings Schichtung vorliegt, so nimmt die Power ab, mit der der TDT eine vorhandene Kopplung entdecken kann (KAPLAN *et al.*, 1997). Gegenüber klassischen Kopplungstests in Form von Affected Sib Pairs ist der TDT bei gleichen Daten dennoch sensitiver (SPIELMAN *et al.*, 1993).

Über die genannten einfachen Familientrios hinaus ist der TDT als Kopplungstest auch in Familienstammbäumen und mit mehreren erkrankten Kindern einer Familie anwendbar (EWENS & SPIELMAN, 1995). Das heißt, die notwendigen Familientrios können aus Familienstammbäumen entnommen werden, und aus einer Familie mit zwei erkrankten Kindern können zwei Trios gebildet werden. Die Form der Vererbung ist für

die Gültigkeit des TDT nicht entscheidend. Unter Berücksichtigung der verschiedenen Vererbungsmodelle lassen sich jedoch verschiedene ähnliche Tests entwickeln, die eine höhere Power besitzen, Kopplung zu erfassen (SCHAID & SOMMER, 1994).

Der klassische TDT ist auch ein Test auf Assoziation, allerdings für unabhängige Familien mit einem Kind (SPIELMAN & EWENS, 1996). MARTIN *et al.* (1997) haben den TDT so modifiziert, so daß er auch bei Verwendung von Familien mit mehreren kranken Kindern ein gültiger Test auf Assoziation bleibt. Mit dem PDT (MARTIN *et al.*, 2000; MARTIN *et al.*, 2001) wurde schließlich ein gültiger Test auf Assoziation vorgestellt, der auch mehrere Familien aus Stammbäumen verwenden darf.

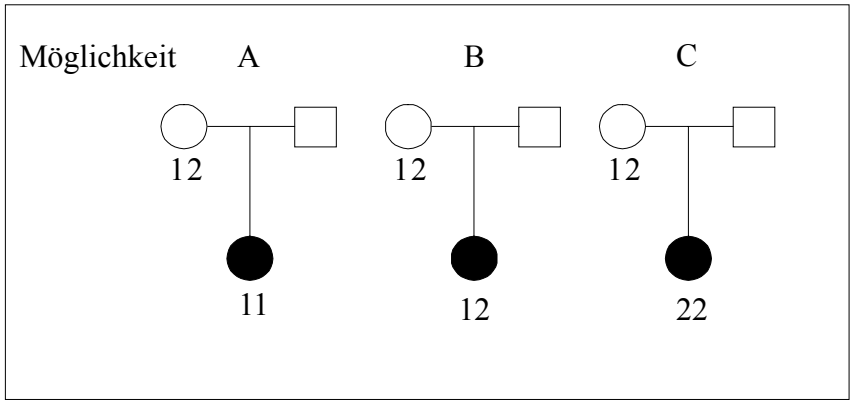
Der TDT ist in den letzten Jahren häufig erfolgreich verwendet worden. Beim Diabetes mellitus Typ I ist es für etwa 10 Genregionen gelungen, mögliche Krankheitsloci auf einen sehr engen Bereich einzuschränken. Als erste haben SPIELMAN *et al.* (1993) mit dem TDT Kopplung zwischen einem Restriktionslängenpolymorphismus (5´FP) aus der Region des Insulingens auf Chromosom 11q und Diabetes mellitus Typ I (IDDM) nachgewiesen. Vorher wurde lediglich Assoziation beschrieben und Kopplungsstudien mit affected-sib pairs (ASP) konnten keinen Beweis für Kopplung liefern. Die Restriktionsfragmente von 5´FP wurde in Abhängigkeit von ihrer Länge in drei Klassen eingeteilt. Das Klasse I Allel wurde in 78 Fällen an das erkrankte Kind vererbt, die beiden anderen Allele nur in 46 Fällen. Der Testwert des TDT betrug χ^2=8,26 und der p-Wert p=0,004. Der Unterschied bei der Vererbung der Allele ist damit signifikant und zeigt Kopplung in Anwesenheit von Assoziation.

6.3 Probleme bei fehlenden Daten

Da der TDT immer auch auf die Genotypen der Eltern angewiesen ist, gibt es Probleme, wenn sich Krankheiten erst im fortgeschrittenen Alter manifestieren. In diesem Fall können häufig nicht mehr die Eltern am Markerlocus typisiert werden, weil sie entweder bereits verstorben sind oder aus anderen Gründen nicht mehr rekrutiert werden können. Mögliche andere Gründe wären, daß ein Elternteil z.B. wegen Scheidung nicht erreichbar ist oder daß sich bei der Typisierung z.B. ein Nicht-Vaterschaft herausstellt. Der Untersucher ist dann vor das Problem gestellt, daß es ihm nicht mehr möglich ist, alle Familientrios vollständig am Markerlocus zu typisieren. Dennoch gibt es auch bei Paaren aus einem Elternteil und einem erkrankten Kind bestimmte Fälle in denen es möglich ist, formal einen Eintrag in die T/NT-Tabelle

vorzunehmen. CURTIS & SHAM (1995) habe jedoch gezeigt, daß ein Bias entsteht, wenn diese Fälle mit in die TDT-Statistik einfließen und die nicht eindeutigen Fälle unberücksichtigt bleiben. In Abbildung 6-4 sind alle drei Möglichkeiten bei einem heterozygoten Elternteil illustriert.

Abbildung 6-4 Mögliche Elter-Kind Paare, wenn ein Elternteil heterozygot ist und die Genotypinformation des anderen Elternteil fehlt



Bei Möglichkeit B mit einem heterozygoten Kind und einem heterozygoten Elternteil kann nicht erkannt werden, welches Allel transmittierte und welches Allel nicht transmittierte wird. Damit ist Möglichkeit B uninformativ und kann nicht berücksichtigt werden. Bei Möglichkeit A und C mit einem homozygotem Kind und einem heterozygoten Elternteil kann eindeutige zugeordnet werden, welches Allel vererbt wurde. In Möglichkeit A wurde Allel 1 vererbt, und in Möglichkeit C wurde Allel 2 vererbt. Ob Möglichkeit A oder C häufiger eintritt, ist allerdings abhängig von der Frequenz P(1) und P(2), mit dem die Allele in der Population vorkommen. Ist also beispielsweise das Allel 1 häufiger, so wird auch Möglichkeit A häufiger auftreten. Damit sind die Einträge nicht mehr alleine anhängig von Assoziation und Kopplung, sondern auch von den Allelfrequenzen in der Population. Es resultiert also ein Bias, wenn dennoch die Informationen von inkompletten Familien berücksichtigt werden.

6.3.1 Geschwister-TDT

Im Zuge der Problematik des klassischen TDT bei fehlenden Informationen von Genotypen der Eltern haben SPIELMAN & EWENS (1998) einen neuen Test, den Sib-TDT, vorgeschlagen. Anstelle der elterlichen Genotypen werden zum Testen auf Kopplung und Assoziation die Genotypen der nicht erkrankten Geschwister benötigt. Wie bei einer gematchten Fall-Kontrollstudie wird verglichen, ob erkrankte Geschwisterkind überzufällig häufig ein spezifisches Markerallel besitzt als nicht

erkrankte Geschwister. Bei dem SDT von HORVATH & LAIRD (1998) werden zuerst Frequenzen eines spezifischen Markerallels unter den gesunden und kranken Geschwistern einer Familie verglichen. Die Teststatistik wird aus der Anzahl von Geschwistergruppen, bei denen kranke Geschwister häufiger das spezifische Allel besitzen, und der Anzahl von Geschwistergruppen, bei denen gesunde Geschwister häufiger das spezifische Allel besitzen, gebildet. Außerdem wurde der SDT auch auf multiallelische Marker erweitert.

Dieses Rekrutierungsschema sollte jedoch in der Praxis stets sorgfältig geprüft werden, da dieses Design bei einem rezessiven Modell nur eine geringe Power besitzt (ZIEGLER & HEBEBRAND, 1998). Zwar scheint der Ansatz bei Geschwisterpaaren mit extremer Diskordanz in Bezug auf den Phänotyp der Erkrankung besonders praktikabel, aber kulturelle und soziale Faktoren können dieses Phänomen verzerren. So fanden ZIEGLER & HEBEBRAND (1998) in einer Studie heraus, daß sich bei einer Jugendlichen mit Anorexia nervosa das extreme Untergewicht vermutlich als Reaktion auf die Adipositas der Schwester entwickelt hatte. Weiterhin kann Diskordanz entstehen, wenn bei altersabhängigen Penetranzen jüngere Geschwister als Kontrollen verwendet werden. Darüber hinaus gibt es Probleme bei Krankheiten mit reduzierten Penetranzen, wie z.B. familiärem Brustkrebs mit den Genen BRCA1 und BRCA2 (FORD *et al.*, 1998) oder Retinopathia pigmentosa (MCGEE *et al*., 1998). Außerdem sind der Sib-TDT und SDT nur anwendbar, wenn bei fehlenden elterlichen Daten Geschwister vorhanden sind. In dem Fall bei dem nur das erkrankte Kind und ein Elternteil typisiert werden können, sind der Sib-TDT und SDT nicht verwendbar.

6.3.2 Rekonstruktions-Tests

KNAPP (1999) stellte für das Problem bei fehlenden Daten einen anderen Test, den sogenannten „reconstruction combined TDT“ (RC-TDT), vor und verglich diesen Test mit dem Sib-TDT und dem klassischen TDT. Die Idee des RC-TDT ist, den fehlenden Genotyp eines Elternteils zu rekonstruieren und den entstehenden Bias zu korrigieren. Als Markerlocus benötigt KNAPP (1999) jedoch multiallelische Marker. Darüber hinaus benötigt der RC-TDT sowohl erkrankte als auch gesunde Geschwisterkinder. Bei einem hohen Maß an Markerpolymorphismus und einer hohen Fallzahl hat der Ansatz des RC-TDT eine höhere Power als der Sib-TDT, da in diesem Fall die Wahrscheinlichkeit steigt, den fehlenden Genotypen aus den Genotypinformationen der Kinder zu bestimmen. Da für den RC-TDT ein multiallelischer Marker und gesunde

Geschwisterkinder notwendig sind, kann er nicht auf die bisher beschriebenen Elter-Kind Paaren mit einem biallelische Marker angewandt werden. Deshalb wird der RC-TDT in dieser Arbeit nicht weiter untersucht.

Ein ähnliches Verfahren der Rekonstruktion der fehlenden Daten wurden auch von MARTIN *et al.* (1998). Die sogenannte Parental Genotype Reconstruction (PRG) basiert auf dem Haplotype Relative Risk (HRR) und rekonstruiert die fehlenden Daten unter Verwendung der Genotypinformation der Geschwister. Nach Schätzung der Allelfrequenzen wird in den Fällen, in denen nicht eindeutig auf die elterlichen Genotypen geschlossen werden kann, unter Verwendung der geschätzten Allelfrequenzen die internen Kontrollen unter der Annahme keiner Assoziation $\delta=0$ gebildet. Diese Annahme wird aufgestellt, um keine Verzerrung bei der Schätzung zu erzeugen und den Fehler 1. Art nicht zu erhöhen. In Simulationen wurde gezeigt, daß der Fehler 1. Art nur sehr gering steigt und trotz der Annahme von $\delta=0$ die Power steigt, mit der Assoziation nachzuweisen ist. Nachteil der Methode bleibt jedoch auch die zusätzlich notwendige Typisierung von Geschwisterkindern.

Von WEINBERG (1999) wurde ein Verfahren zur Schätzung der fehlenden Genotypen auf der Basis der Maximum-Likelihood-Methode vorgestellt. Dem sogenannten likelihood-ratio test (LRT) liegt wie dem TDT das klassische Triodesign zugrunde. Bei dem Schätzprinzip der Maximum-Likelihood-Methode gilt diejenige Größe als guter Schätzer, die das realisierte Studienergebnis am wahrscheinlichsten macht. Die Schätzung erfolgt unter Verwendung des expectation maximization (EM) Algorithmus. Ein Voraussetzung ist eine Vorstellung über die Verteilung der betrachteten Variablen. Der LRT basiert auf einem log-linearen Modell. In Simulationen wird gezeigt, daß der EM-LRT verlorene Power zurückgewinnt und das nominelle Signifikanzniveau von 5% hält.

6.3.3 1-TDT

Alle bisher vorgestellten Tests sind nicht anwendbar auf die Situation, in der ausschließlich die Informationen für einen biallelischen Marker von einem erkrankten Kind und einem Elternteil vorhanden sind. Wenn die Informationen von einem heterozygoten Elternteil und einem homozygoten Kind für die TDT-Statistik verwendet werden, resultiert, wie bereits zuvor erläutert, ein Bias (CURTIS & SHAM, 1995). Genau für diese Situation wurde von SUN *et al.* (1999) ein sehr interessanter und einfacher

Ansatz vorgeschlagen, der die Verwendung der Informationen ermöglicht, ohne daß daraus ein Bias entsteht.

Die Idee der neuen Teststatistik, dem 1-TDT, ist die Anwendung auf die Situation, wenn nur Informationen über Genotypen an einem biallelische Marker von Elter-Kind-Paaren vorhanden sind. SUN *et al.* (1998) haben einen neuen nicht-iterativen Schätzer für das Odds Ratio in internen Fall-Kontroll Studien gefunden, der auch auf die Situation von Elter-Kind-Paaren übertragbar ist. Gemäß dieses Ansatzes wurde ein genauer und nicht iterativer Schätzer λ für das Odds Ratio zwischen Personen mit den Genotypen 12 und 11 am Markerlocus (λ_1) sowie den Genotypen 12 und 22 am Markerlocus (λ_{-1}) gefunden, wenn nur ein Elternteil vorhanden ist (SUN *et al.*, 1999). Der Schätzer für das OR bleibt annähernd genau, solange die beiden folgenden Annahmen bestehen.

Erste Annahme: Väterliche und mütterliche Genotypen am Markerlocus sind gleich verteilt, d.h. es besteht Hardy-Weinberg Gleichgewicht.

Zweite Annahme: Vater oder Mutter in jeder Kernfamilie fehlen mit der gleichen Wahrscheinlichkeit von ½.

Liegt keine Assoziation zwischen Markerlocus und der Krankheit vor, so sind $\lambda_1=\lambda_{-1}=1$. Bei positiver Assoziation zwischen Markerallel 1und der Krankheit wird $\lambda_1>1$ und/oder $\lambda_{-1}<1$. Aus den beiden Schätzers abgeleitet, wurden in der Arbeit für den 1-TDT zwei Teststatistiken vorgeschlagen. Beide Statistiken können sowohl in familienbasierten Assoziationsstudien als auch in Kopplungsstudien verwendet werden, ohne einen Bias unter den Bedingungen der Nullhypothese mit keiner Assoziation oder Kopplung zu erzeugen. Die erste Statistik T_1 ist anwendbar, wenn eine der beiden folgenden Annahmen erfüllt sind. Die Statistik T_2 ist ein gültiger Test, wenn beide Annahmen verletzt sind. SUN *et al.* (1999) haben theoretisch und in Simulationen gezeigt, daß der 1-TDT das nominale Signifikanzniveau von 5% hält und zu einem Powergewinn führt.

Im folgenden wird die Berechnung des 1-TDT erläutert. Für den 1-TDT wurde eine (3x3)-Tabelle definiert, die den Genotypen am Markerlocus des Kindes und des vorhandenen Elternteils enthält.

Tabelle 6-5 Fall-Kontroll Design, wenn nur ein Elternteil vorhanden ist

Genotyp des erkrankten Kindes	**Genotyp des vorhandenen Elternteils**		
	11	12	22
11	A_{00}	$\mathbf{A_{01}}$	0
12	$\mathbf{A_{12}}$	A_{11}	$\mathbf{A_{12}}$
22	0	$\mathbf{A_{21}}$	A_{22}

Ähnlich dem klassischen TDT testet auch der 1-TDT auf Homogenität in der Tabelle in Bezug auf die Nebendiagonale. Von den Schätzern λ_1 und λ_{-1} für das Risikoverhältnis abgeleitet, ergeben sich die Variablen b_1 und c_1, die miteinander verglichen werden.

$$b_1=A_{01}+A_{12} \text{ und } c_1=A_{10}+A_{21}$$

Wenn die Erkrankung mit dem Markerallel 1 assoziiert ist, ist $b_1-c_1 \neq 0$. Die Varianz von b_1-c_1 wird geschätzt durch $V= b_1+c_1$. Daraus ergibt sich die folgende Statistik, wenn nur ein erkranktes Kind und ein vorhandener Elternteil berücksichtigt werden.

$$T_1 = \frac{b_1 - c_1}{\sqrt{V}}$$

T_1 folgt unter der Nullhypothese „keine Assoziation oder keine Kopplung“ asymptotisch einer Standardnormalverteilung. Entsprechend ist

$$T_1^{\,2} = \frac{(b_1 - c_1)^2}{b_1 + c_1}$$

unter der Nullhypothese asymptotisch χ^2-verteilt mit 1 Freiheitsgrad.

In der Praxis liegen in der Regel unterschiedliche Typen von Familien vor. Es ist daher wünschenswert, die Statistiken des 1-TDT, des Sib-TDT und des klassischen TDT verbinden zu können.

Dieses ist aufgrund der Unabhängigkeit der Familien und den asyptomatischen Eigenschaften der Teststatistiken möglich. Im Rahmen dieser Arbeit wird die Kombination des klassischen TDT mit dem 1-TDT beschrieben.

Es seien b und c aus Kapitel 6.1 die relativen Einträge der klassischen TDT-Tafel. Dann ist die unter der Nullhypothese erwartete Anzahl der Einträge in Feld b gleich $\frac{b+c}{2}$.

Die Varianz ist ebenfalls $\frac{b+c}{2}$. Mit den Bezeichnungen aus diesem Abschnitt ergibt sich analog für den 1-TDT als Erwartungswert $\frac{b_1+c_1}{2}$ sowie als Varianz $\frac{b_1+c_1}{4}$.

Damit folgt die Statistik W=b+b1 unter H_0 asymptotisch einer Normalverteilung mit dem Mittelwert $\frac{b+c}{2}+\frac{b_1+c_1}{2}$ und der Varianz $\frac{b+c}{2}+\frac{b_1+c_1}{4}$. Entsprechend ist die Größe

$$Z=\frac{W-\frac{b+c}{2}-\frac{b_1+c_1}{2}}{\sqrt{\frac{b+c}{2}-\frac{b_1+c_1}{4}}}$$

unter H_0 asymptotisch standardnormalverteilt.

SUN *et al.* (1999) haben den 1-TDT auf den gleichen Datensatz von Patienten mit IDDM und dem VNTR Locus am 5´Ende des Insulingens (5´FP) wie in SPIELMAN *et al.* (1993) und SPIELMAN & EWENS (1998) angewandt. Für das Klasse I Allel wurde bei den Fällen mit dem Vater und einem erkrankten Kind der Testwert χ^2=3,13 und bei den Fällen mit einer Mutter und einem erkrankten Kind der Testwert χ^2=2,28 berechnet. In beiden Fällen konnte damit bei den Elter-Kind Paaren keine Assoziation und Kopplung nachgewiesen werden.

Anmerkung: In der Arbeit von SUN *et al.* (1999) wird auf Seite 100 M_1 ursprünglich als $M_1 = (b_1 + c_1)/2$ definiert. Setzt man dieses M_1 jedoch so in die Gleichung von A_{com} ein, wird A_{com} zu klein. Unter den Bedingungen der Nullhypothese wäre dann der Term ($W-A_{com}$) nicht mehr nahe 0. Daher wird M_1 nicht als $M_1 = b_1 + c_1$ definiert. Das diese Korrektur zulässig ist, kann in meinen Simulationen nachvollzogen werden, wenn man sich eine Matrix von Tabelle 6-5 unter der Nullhypothese ausgeben läßt.

7 Rekonstruktionen mit dem EM-Algorithmus

7.1 Ansatz der Rekonstruktion

In dieser Arbeit wird der neue Gedanke verfolgt, ob es möglich und praktikabel ist, mit Hilfe der verbleibenden Daten aus den kompletten Familientrios und den Paaren Rückschlüsse auf den unbekannten Genotyp des fehlenden Elternteils zu ziehen. Mit den vorhandenen Daten kann unter Verwendung des EM-Algorithmus abgeschätzt werden, welcher der möglichen Genotypen am ehesten auf den fehlenden Elternteil zutrifft. Zu den oben genannten Überlegungen wurde ein Computerprogramm mit folgenden Prozeduren geschrieben. In der ersten Prozedur können beliebig viele geeignete Familientrios generiert werden. Die Generierung der Familien erfolgt aus zwei verschiedenen Populationen. Der Anteil der Populationen an der Gesamtpopulation wird je nach Simulationsmodell variiert. Unter Vorgabe des Vererbungsmodells, der Genfrequenzen des Krankheitsallels und des Markerallels, des Assoziationsparameters und der Rekombinationsrate ist es möglich, Simulationen sowohl unter den Bedingungen der Nullhypothese als auch der Alternativhypothese durchzuführen. In der zweiten Prozedur wird bei einem Teil der Familientrios jeweils ein Elternteil zufällig gelöscht. Die dritte Prozedur rekonstruiert. die fehlenden Daten mit dem EM-Rekonstruktion anhand des folgenden Algorithmus.

1. Aus den Eltern der vollständigen Familien werden die Genotyp- bzw. Allelfrequenzen geschätzt.
2. Schätzung der Genotypen der fehlenden Elternteile unter Verwendung der geschätzten Genotyp- bzw. Allelfrequenzen.
3. Neuschätzung der Genotyp- bzw. Allelfrequenzen unter Verwendung aller elterlicher Daten, einschließlich der rekonstruierten Elternteile.
4. Wiederholung der Schritte 2 und 3, bis die Genotyp- bzw. Allelfrequenzen hinreichend stabil geschätzt sind ($\varepsilon > 10^{-5}$).
5. Bei jedem Elter-Kind Paare wird danach geprüft, werde Genotypen für den fehlenden Elternteil möglich sind. Bei der EM-Allelrekonstruktion werden die Häufigkeiten der möglichen Genotypen unter Verwendung der geschätzten Allelfrequenzen berechnet, bei der EM-Genotyprekonstruktion werden die geschätzten Genotypfrequenzen zugrundegelegt.

6. Der Eintrag in die T/NT-Tabelle erfolgt dann relative zu den entsprechenden Genotypwahrscheinlichkeiten. Dieses Vorgehen ist möglich, da für den Eintrag in die T/NT-Tabelle und für die Berechnung der TDT-Statistik ganze Zahlen nicht zwingend notwendig sind.

Außerdem wird der 1-TDT in das Programm implementiert. In getrennte T/NT-Tabellen werden anschließend alle Familientrios, die vollständigen Familientrios nach Löschen von Genotypinformationen und die Familientrios nach Rekonstruktion der fehlender Daten eingetragen. Für alle T/NT-Tabellen wird abschließend die TDT-Statistik berechnet.

Das Programm ist in der Programmiersprache GAUSS© (Version 3.2.28, 1997) geschrieben. Die Auswahl der Programmiersprache basiert auf den Überlegungen von SCHEPERS (1991, S.61f.) und ENACHE (1994, S. 19f.). Die generierten Familiendaten werden sinnvollerweise in Form von Matrizen dargestellt. Dazu ist eine Programmiersprache notwendig, die einfache Matrixoperationen und Matrixmanipulationen erlaubt. GAUSS© ist als strukturierte Hochsprache für mathematisch-statistische Anwendung konzipiert und basiert im Kern auf dem Datentyp einer Matrix von Gleitkommazahlen. Das ermöglicht die direkte Matrizenschreibweise von üblichen Berechnungen. Außerdem verfügt GAUSS© über vielseitige Strukturierungsmöglichkeiten von konditionalen Anweisungen, Schleifen usw. sowie flexiblen Prozedurkonzepten. Das erlaubt die Entwicklung von modularen und gut strukturierten Programmen. Als Zufallsgenerator wird die GAUSS©-Prozedur `rndu` verwendet, die einen linear kongruenten Algorithmus benutzt. Als Seed wird die Systemzeit gewählt, für Details siehe KENNEDY & GENTLE (1980).

7.2 Rekonstruktion der fehlenden Daten

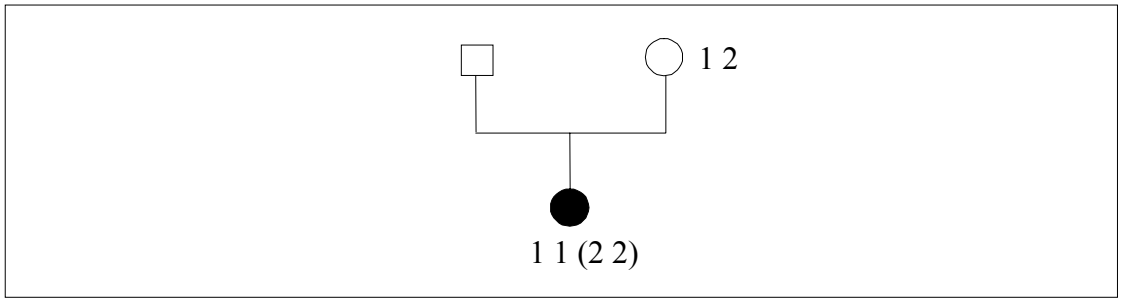
Die Paare aus erkranktem Kind und Mutter sind in ihrem Informationsgehalt nicht einheitlich. Abhängig von der Konstellation der bekannten Allele zueinander lassen sie sich in zwei Gruppen unterteilen. In der einen Konstellation kann auf das vererbte Allel des Vaters eindeutig geschlossen werden, so daß nur das nicht vererbte Allel unbekannt bleibt. In der anderen Konstellation kann der Vater beide Allelvarianten vererbt haben. In diesem Fall sind beide Allele des Vaters unbekannt.

Die Idee, die fehlenden Daten möglichst genau abzuschätzen, kann auf verschiedene Weise umgesetzt werden. Eine Möglichkeit ist, die unbekannten Allele des Vaters

einzeln zu rekonstruieren. Ist das vom Vater vererbte Allel bekannt, muß nur das nicht vererbte Allel geschätzt werden. Sind beide Allele jedoch unbekannt, werden beide Allele geschätzt. Dieses Vorgehen wird im weiteren EM-Allelrekonstruktion genannt. Dieses Vorgehen berücksichtigt allerdings nicht Beziehungen zwischen den beiden Allelen, wenn die Population nicht in Hardy-Weinberg Gleichgewicht ist. Daher wird als alternatives Verfahren bei der EM-Genotyprekonstruktion der Genotyp am Markerlocus des unbekannten Vaters geschätzt. Das heißt dieses Vorgehen berücksichtigt die Häufigkeiten der Kombinationen beider Allele.

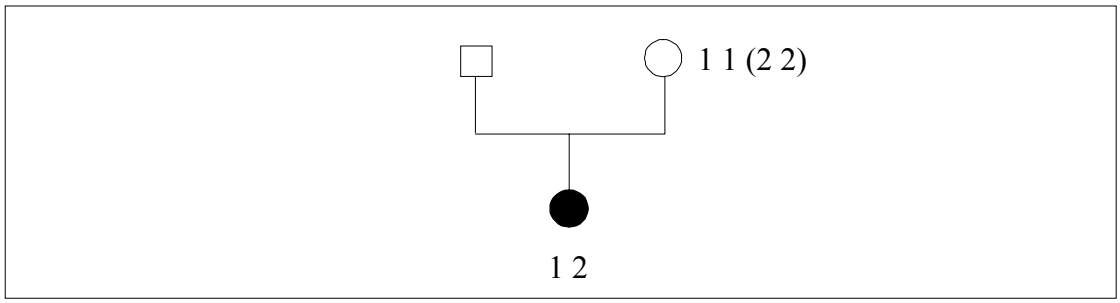
Sind Mutter und Kind nicht gleichzeitig heterozygot, gibt es zwei mögliche Varianten der Genotypverteilungen, bei denen allein die Vererbung auf väterlicher Seite unbekannt ist. In der ersten Varianten ist die Mutter heterozygot und das Kind homozygot, in der zweiten Variante ist die Mutter homozygot und das Kind heterozygot. Bei beiden Varianten lassen sich die transmittierten und nicht transmittierten Allele der Mutter klar zuordnen.

Abbildung 7-1 Die Mutter ist heterozygot und das Kind ist homozygot, transmittiertes und nicht transmittiertes Allel der Mutter lassen sich klar zuordnen



Hat das Kind den Genotyp 11, so besitzt der Vater entweder den Genotypen 11 oder 12 am Markergenlocus. Hat das Kind den Genotyp 22, so besitzt der Vater entweder den Genotypen 12 oder 22 am Markergenlocus.

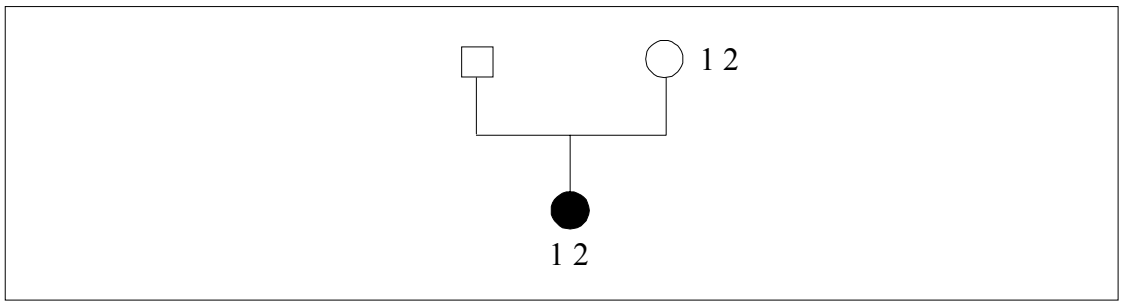
Abbildung 7-2 Die Mutter ist homozygot und das Kind ist heterozygot, transmittiertes und nicht transmittiertes Allel der Mutter lassen sich klar zuordnen



Ist das Kind heterozygot 12 und die Mutter homozygot 11, dann hat der Vater entweder den Genotyp 12 oder 22. Ist das Kind heterozygot 12 und die Mutter 22, dann hat der Vater den Genotyp 12 oder 11.

Im zweiten Fall sind Mutter und das erkrankte Kind beide heterozygot am Markerlocus. In diesem Fall kann für die Mutter nicht entschieden werden, ob sie das Allel 1 oder 2 an das Kind transmittiert. Für diese Entscheidung wäre die Kenntnis des väterlichen Genotyps notwendig.

Abbildung 7-3 Mutter und Kind sind beide heterozygot, transmittiertes und nicht transmittiertes Allel der Mutter lassen sich nicht klar zuordnen



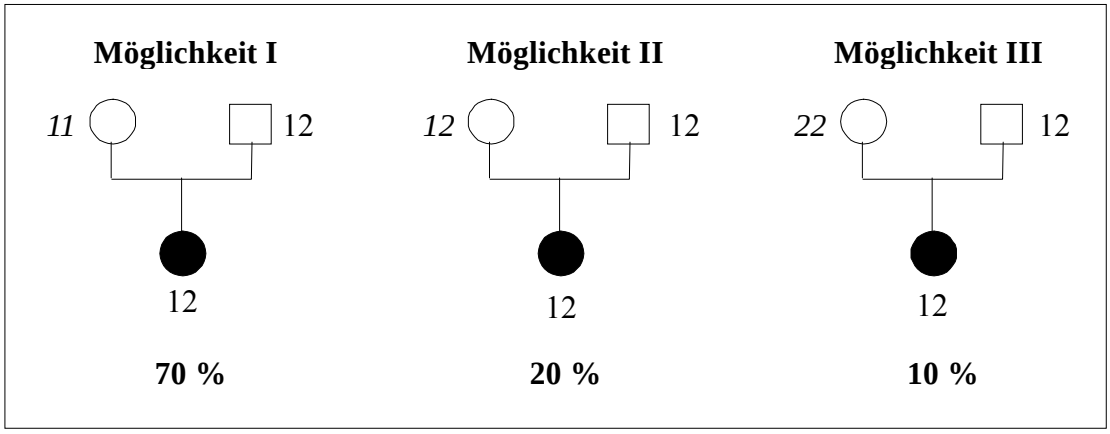
Hier kann aufgrund fehlender weiterer Informationen nur eine Wahrscheinlichkeit von 0,5 für die Transmission der mütterlichen Allele angenommen werden. Bei eventuell vorliegender Assoziation wäre der wahre Werte jedoch stark von 0,5 verschieden, so daß bei diesem Vorgehen ein zusätzlicher Fehler erzeugt würde. Für die väterlichen Genotypen sind alle Varianten 11, 12 und 22 möglich.

7.2.1 Rekonstruktion nach der größten Wahrscheinlichkeit

Anhand der geschätzten Genotyp- bzw. Allelfrequenzen ergeben sich für die möglichen väterlichen Genotypen bestimmte Wahrscheinlichkeiten. Bei der Entscheidung, welcher Genotyp rekonstruiert werden soll, liegt intuitiv am nächsten, sich für den wahrscheinlichsten Genotypen zu entscheiden. Die Prozeduren `REKONSTR2` und

REKONSTR5 verfahren nach diesem Prinzip. Dieses Verfahren mißachtet allerdings alle weniger wahrscheinlichen Möglichkeit. Ein Beispiel ist in Abbildung 7-4 dargestellt. Bei einer heterozygoten Mutter und einem heterozygoten Kind gibt es für den väterlichen Genotypen die Möglichkeiten I, II und III mit den Wahrscheinlichkeiten 70, 20 und 10%

Abbildung 7-4 Beispiel für drei mögliche Genotypen für den Vater (*kursiv*) und entsprechende Wahrscheinlichkeiten, wenn die Mutter und das Kind heterozygot sind.



Wie beschrieben, würde man sich jetzt für Möglichkeit I mit der Wahrscheinlichkeit von 70% entscheiden. Die Möglichkeiten II und III, zusammen 30% der möglichen Genotypen, werden dann allerdings nicht berücksichtigt. Offensichtlich erzeugt dieses Verfahren dadurch einen Bias. Daher wird dieses Vorgehen nicht weiter verfolgt.

7.2.2 Rekonstruktion nach Zufallsereignissen

Weiterhin wäre auch möglich, bei jeder Rekonstruktion eines väterlichen Genotyps eine Zufallszahl z im Intervall [0;1] zu werfen, und danach zu entscheiden, für welchen Genotyp man sich entscheidet. Entsprechend dem Beispiel aus Abbildung 7-4 entscheidet man sich für Möglichkeit I, wenn $0 \leq z \leq 0{,}7$, für Möglichkeit II, wenn $0{,}7 < z \leq 0{,}9$, und für Möglichkeit III, wenn $0{,}9 < z \leq 1$ In den Prozeduren REKONSTR3 und REKONSTR6. Es ist zu erwarten, daß auf lange Sicht bei einer ausreichenden Zahl an rekonstruierten Genotypen dieses Vorgehen alle Möglichkeiten ausreichend repräsentativ abbildet. Bei einer kleinen Anzahl untersuchter Familien ist dieses Verfahren allerdings fehleranfällig und starken Zufallsschwankungen unterworfen. Daher wird dieses Vorgehen auch nicht weiter verfolgt.

7.2.3 Rekonstruktion gemäß der Genotypwahrscheinlichkeiten

Wenn für den Vater zwei oder drei Möglichkeiten bestehen, wäre es auch denkbar, möglichen Genotypen relativ zu ihren Wahrscheinlichkeit zu berücksichtigen. Dieses Vorgehen ist möglich, weil für den Eintrag in die T/NT-Tabelle und für die Berechnung des Testwertes ganze Zahlen nicht zwingend notwendig sind. Die TDT-Statistik kann für alle reellen Zahlen berechnet werden. Tabelle 7-1 zeigt, wie der Eintrag der väterlichen Transmission in die T/NT-Tabelle für das Beispiel aus Abbildung 7-4 erfolgt.

Tabelle 7-1 Beispiel für den relativen Eintrag der väterlichen Transmission in die T/NT-Tabelle bei drei möglichen Genotypen entsprechend der Wahrscheinlichkeiten

Möglichkeit I

T	NT	
	1	2
1	|	
2	|	

T	NT	
	1	2
1	0,7	
2	0,7	

Möglichkeit II

T	NT	
	1	2
1		|
2	|	

T	NT	
	1	2
1		0,2
2	0,2	

Möglichkeit III

T	NT	
	1	2
1		|
2		|

T	NT	
	1	2
1		0,1
2		0,1

Die Addition der Einträge aus allen Zellen ergibt die Summe 2, genau identisch mit einem Eintrag einer echten Transmission. Diese Rekonstruktion erfolgt mit den Prozeduren `FAMCREA1` und `FAMCREA4`. In der Summe geben alle Einträge immer zwei, wie bei Einträgen mit ganzen Zahlen.

7.2.4 EM-Allelrekonstruktion

Um eine Rekonstruktion auf der Basis der Allele durchzuführen, müssen aus den vorhandenen Daten die notwendigen Wahrscheinlichkeiten beiden Allele geschätzt werden. Für die Verteilung der Allele 1 und 2 am Markerlocus wird eine Hardy-Weinberg Verteilung angenommen. Danach ergeben sich für die Genotypen der Eltern die Wahrscheinlichkeiten

$$P(11) = m^2$$
$$P(12 \text{ oder } 21) = 2m\ (1-m)$$
$$P(22) = (1-m)^2$$

Im folgenden wird die Matrize `ergm` für die Anzahl von (`n-v`) vollständigen Familientrios und die Matrize `pairs` für die Anzahl von `v` Paare verwendet. Der genaue Aufbau der Matrizen wird später erläutert.

Als Beispiel wird angenommen, daß bei einer Studie 80 Familientrios mit vollständigen Genotypen und 20 Paare mit fehlendem väterlichen Genotyp bekannt sind. Die erste Frage ist jetzt: „Welcher Genotyp wird bei dem fehlenden Elternteil erwartet?“. Zur Abschätzung der Genfrequenz `em1` ist in der Statistik die Maximierung durch den klassischen EM-Algorithmus („expectation maximization algorithm“) (LAIRD, 1993) üblich. Dieser Vorgehen ist in der Prozedur `ESTIMm1` implementiert worden:

```
{em1} = ESTIMm1 (ergm, pairs, n, v);
```

Dazu werden aus den vollständigen Trios in der Matrix `ergm` die Markerallelfrequenzen mit dem EM-Algorithmus geschätzt. Als Beispiel lassen sich aus 80 Trios die Genfrequenzen für das Allel 1 P(1) = `m` und für das Allel 2 P(2) = (`1-m`) bestimmen. Wenn eine Hardy-Weinberg Gleichgewicht vorliegt, so läßt sich die Frequenz der Allele über die Maximum-Likelihoodmethode für eine Binomialverteilung schätzen

Die Variable (`n-v`) ist die Anzahl aller Allele bei den Eltern der vollständigen Trios. Da jedes Trio zwei Elternteile mit zwei Allelen enthält, ist in dem Beispiel (`n-v`) = 320. Die Variable `k` gibt die Anzahl der Allele 1 unter den 320 Allelen an. Bekannt sind demnach nur `n` und `k`. Die Wahrscheinlichkeit m wird so geschätzt, daß es bei einem bestimmten (`n-v`) und `k` am wahrscheinlichsten ist, das ist `em1`=`k`/(`n-v`). Aus den 80 Trios läßt sich auf diese Weise `em1` aus den Genotypen der Eltern schätzen.

Diese Allelfrequenzen werden dann auch für die fehlenden Väter bei den Paaren aus der Matrix `pairs` angenommen. Im nächsten Schritt werden dann wiederum die Allelfrequenzen aus den vollständigen Trios und den ergänzten Trios neu geschätzt. Falls die Differenz zwischen der zuvor geschätzten Frequenz und der aktuell geschätzten Frequenz kleiner als $\varepsilon = 10^{-5}$ ist, wird das Rekonstruktionsverfahren beendet. Ist die Frequenzdifferenz ausreichend gering, wird angenommen, daß die

Allelfrequenz stabil ist. Ist die Frequenzdifferenz jedoch nicht größer als $\varepsilon = 10^{-5}$ wird dieses Verfahren so lang wiederholt, bis die Frequenzdifferenz ausreichend gering ist. Am Ende wird die geschätzte Allelfrequenz `em1` aus der Prozedur zurückgeben.

Eine andere Möglichkeit wird in der Prozedur `ESTIMm2` angewandt.

```
{em2} = ESTIMm2 (ergm, pairs, n, v)
```

Die Allelfrequenz `em2` des Markerallels wird aus den Daten aller Eltern geschätzt. Das heißt, es werden in dem Beispiel zugleich die 80 Trios und die 20 Paare für die Schätzung verwendet. Da sich die nachfolgende Rekonstruktion auf die Gesamtheit der verfügbaren Daten bezieht, wird die Allelfrequenz auch mit den neu gewonnenen Elternteilen konstant bleiben.

Die Endergebnisse der geschätzten Markerallelfrequenz sollte in beiden Ansätzen identisch sein. Der Programmtest (vgl. Kapitel 7.5.2) zeigt die Gleichwertigkeit beider Ansätze in Bezug auf das Endergebnis. Da in der Prozedur `ESTIMm2` das Ergebnis in nur einer Schleife berechnet wird, wird aufgrund des Geschwindigkeitsvorteils im weiteren nur die Prozedur `ESTIMm2` verwendet.

Die zweite Frage stellt sich jetzt: „Welcher Genotyp wird mit Hilfe der gewonnenen Information für den fehlenden Elternteil angenommen?“. Dazu wurde die Prozedur `REKONSTR1` geschrieben, die die Rekonstruktion gemäß der Genotypwahrscheinlichkeiten (vgl. Kapitel 7.2.1) durchführt.

```
{tntr1} = REKONSTR1 (pairs, v, em2)
```

Für den zuerst beschriebenen Fall, bei dem die Genotypen von Mutter und Kind nicht gleichzeitig heterozygot sind, läßt sich das transmittierte und nicht transmittierte Allel der Mutter bestimmen. Daraus folgt, daß das zweite kindliche Allel vom Vater stammen muß. Ein Allel des Vaters ist damit sicher bekannt. Deshalb muß lediglich das zweite väterliche Allele rekonstruiert werden. Da nicht nur die Genotypen der Eltern, sondern auch die Allele eines Individuums unter der Nullhypothese und Hardy-Weinberg Gleichgewicht voneinander unabhängig sind, kann das fehlende Allel mit Hilfe von `em2` rekonstruiert werden.

Beim zweiten und schwierigeren Fall sind Mutter und Kind gleichzeitig heterozygot. Der Vater kann alle drei Kombinationen am Markerlocus besitzen. Mit Hilfe der Hardy-Weinberg Gleichungen und dem geschätzten `em2` lassen sich die Wahrscheinlichkeiten

für die Genotypen am Markerlocus berechnen. Die Wahrscheinlichkeiten sind auf den Vater ohne Einschränkungen übertragbar, weil bei Vorliegen des Hardy-Weinberg Gleichgewichts die Genotypen der Eltern unabhängig voneinander sind. In dem Beispiel wurden folgende Genotypwahrscheinlichkeiten angenommen.

$$P(1,1) = 0{,}7;\ P(1,2) = 0{,}2;\ P(2,2) = 0{,}1.$$

Die rekonstruierten Trios werden anschließend relativ zu ihrer Wahrscheinlichkeit in die Matrix `tntr1` eingetragen (vgl.Tabelle 7-1) und am Ende der Prozedur an das Hauptprogramm zurückgegeben. Im Hauptprogramm wird nach Kombination der Matrizen `tntm` und `tntr1` unter Verwendung der Prozedur `VALUE` der Testwert der TDT-Statistik berechnet.

7.2.5 EM-Genotyprekonstruktion

Um eine Rekonstruktion auf der Basis der Genotypen durchzuführen, müssen wiederum aus den vorhandenen Daten die notwendigen Wahrscheinlichkeiten bestimmter Genotypen geschätzt werden. Bei der Schätzung dieser Genotypen wird das gleiche Schätzverfahren wie bei der Schätzung der Allelfrequenz verwendet, d.h. bei der Schätzung werden sofort alle vorhandenen Daten der Eltern verwendet. Ein Programmtest (vgl. Kapitel 7.5.2) zeigt auch hier Gleichwertigkeit und einen Geschwindigkeitsvorteil gegenüber dem klassischen Ansatz.

Im Gegensatz zur Rekonstruktion der Allele, bei der ein einziger Parameter geschätzt wird, sind bei der Schätzung für die Genotyprekonstruktion elf unterschiedliche Wahrscheinlichkeiten abhängig von den Genotypen des Kindes und der Mutter zu bestimmen. Ansonsten ist das Vorgehen bei der Schätzung in Prozedur `ESTIMg2` identisch mit dem Vorgehen in Prozedur `ESTIMm2`.

```
{rg2} = ESTIMg2 (ergm, pairs, n, v);
```

Für den Fall, daß Kind und Mutter heterozygot am Markerlocus sind, kann der Vater alle drei verschiedenen Genotypen besitzen. Aus den beiden Elternteilen der vollständigen Familien und dem Elternteil der unvollständigen Familien werden die drei Wahrscheinlichkeiten P(1,1), P(1,2 oder 2,1) und P(2,2) geschätzt. Daraus werden für den Vater folgende Wahrscheinlichkeiten bestimmt.

P(Vater 11 | Mutter 12, Kind 12, Kind krank) ≙ `rg2`[1,1]

P(Vater 12 | Mutter 12, Kind 12, Kind krank) ≙ `rg2`[2,1]

P(Vater 22 | Mutter 12, Kind 12, Kind krank) ≙ `rg2`[3,1]

Sind die Genotypen der Mutter und des Kindes nicht gleichzeitig heterozygot, ist ein Allel des Vaters sicher. Damit ergeben sich pro Fall für den Vater zwei verschiedene Genotypen.

Ist das Kind homozygot für das Markerallel 1, sind für den Vater die Wahrscheinlichkeiten

P(Vater 11 | Kind 11, Kind krank) ≙ `rg2`[4,1]

P(Vater 12 | Kind 11, Kind krank) ≙ `rg2`[5,1]

zu bestimmen.

Ist das Kind heterozygot am Markerlocus, ist zusätzlich der Genotyp der Mutter entscheidend, weil daraus das vererbte Allel des Vaters bestimmt wird.

Ist die Mutter am Markerlocus homozygot für das Markerallel, gilt für den Vater

P(11 | Mutter 22, Kind 12, Kind krank) ≙ `rg2`[6,1]

P(12 | Mutter 22, Kind 12, Kind krank) ≙ `rg2`[7,1]

P(12 | Mutter 11, Kind 12, Kind krank) ≙ `rg2`[8,1]

P(22 | Mutter 11, Kind 12, Kind krank) ≙ `rg2`[9,1]

Ist das Kind homozygot für das alternative Allel 2, ergibt sich für den Vater

P(Vater 12 | Kind 22, Kind krank) ≙ `rg2`[10,1]

P(Vater 22 | Kind 22, Kind krank) ≙ `rg2`[11,1]

Die Prozeduren `REKONSTR4` führen die Rekonstruktion auf der Basis der Genotypen durch.

```
{tntr1} = REKONSTR4 (pairs, v, rg2);
```

Ergebnis der Rekonstruktion ist wie in der Prozedur `REKONSTR1` die Matrix `tntr4`. Analog werden auch hier relative Einträge vorgenommen. Im Hauptprogramm wird nach Kombination der Matrizen `tntm` und `tntr4` unter Verwendung der Prozedur `VALUE` der Testwert der TDT-Statistik berechnet.

7.2.6 1-TDT

In der Prozedur `REKONSTR7` wurde der von SUN *et al.* (1999) vorgeschlagenen kombinierten Test aus 1-TDT und TDT ausprogrammiert (vgl. Kapitel 6.3.3).

```
{z, t1} = REKONSTR7 (tntm, pairs, v)
```

Die Matrix `tntm` entspricht der T/NT-Tabelle für die vollständigen Familientrios und die Matrix `pairs` enthält die Genotypinformationen der Paare. Die verschiedenen Genotypkonstellationen für ein erkranktes Kind und die Mutter aus der Matrix `pairs` werden zuerst in eine (3x3)-Matrix `mat`, die Tabelle 6-5 entspricht, eingetragen. Zur Berechnung der Teststatistik `t1` des 1-TDT werden folgende Variablen definiert. Unter der Nullhypothese folgt für `t1` asymptotisch einer χ^2-Verteilung.

```
b1 =mat [1,2] +mat[2,3]
c1 =mat [1,2] +mat[2,3]
v1 = b1 + c1
```

$$t1 = ((b1 - c1)/\sqrt{v1})^2$$

Zur Berechnung der Teststatistik `z` des kombinierten Tests werden

```
m1   = (b1+c1)
w    = tntm[1,2] + b1
acom = (tnt[1,2] + tnt[2,1])/2) + m1/2
vcom = (tnt[1,2] + tnt[2,1])/4) + v1/4
```

$$z = ((w - acom)/\sqrt{vcom})^2$$

berechnet. Der Testwert `z` folgt unter der Nullhypothese ebenfalls asymptotischen einer χ^2-Verteilung.

7.3 Familiengenerierung

Um Monte-Carlo Simulationen zum Vergleich der verschiedenen Ansätze durchzuführen, sind Daten von geeignete Familientrios notwendig. Die Prozeduren `FAMCREA1` und `FAMCREA2` generieren unter Vorgabe variabler Populationsparameter Familientrios mit einem erkrankten Kind. Der Genotyp am Markerlocus des erkrankten Kindes wird in paternalen und maternalen Haplotypen unterschieden. Zusätzlich liefert das Programm die paternalen und maternalen Haplotypen, die nicht an das Kind transmittiert wurden. Diese Daten werden

anschließend in die Vierfeldertafel für transmittierte und nicht transmittierte Allele eingetragen, anhand derer die TDT-Teststatistik berechnet wird.

Beide Prozeduren erzeugen die notwendigen Daten jedoch auf unterschiedliche Weise. Die Prozedur `FAMCREA1` beginnt jeweils mit der zufälligen Generierung von zwei Elternteilen. Anschließend erfolgt nach den Mendelschen Regeln die zufällige Vererbung der Allele am Krankheitslocus und am Markerlocus zum Kind. Je nach Vererbungsmodell wird dann entschieden, ob das Kind erkrankt oder nicht erkrankt ist. Beide Elternteile werden zufällig aus der Gesamtpopulation gebildet, so daß die Eltern bei einer geringen Häufigkeit des Krankheitsallels selten ein erkranktes Kind erhalten. Deshalb muß eine zufällige Generierung und Vererbung so oft wiederholt werden, bis die geforderte Anzahl von Familien mit einem erkrankten Kind erreicht ist. Die Prozedur `FAMCREA2` benutzt dagegen die Tabelle 3 auf Seite 1089 aus der Arbeit von KNAPP *et al.* (1993). Diese Arbeit enthält bedingte Wahrscheinlichkeiten möglicher elterlicher Genotypen für ein erkranktes Kind. Die Verwendung dieser Wahrscheinlichkeiten erlaubt die direkte Generierung einer geeigneten Familie. Die Vermeidung von Generierungen gesunder und damit uninformativer Familientrios verspricht eine deutliche Beschleunigung der später durchzuführenden Simulationen.

7.3.1 Familiengenerierung mit `FAMCREA1`

Ausgangspunkt für die Simulationen ist ein binärer Krankheitslocus mit den Allelen D und N sowie ein binärer Markerlocus mit den Allelen 1 und 2. Die Frequenz der Allele D und N sei P(D) = `p` und P(N) = `1-p`, die Häufigkeit der Allele 1 und 2 sei P(1) = `m` und P(2) = `1-m`. Die Rekombinationsrate zwischen Krankheits- und Markerlocus wird mit `theta` bezeichnet. Die Penetranzen seien `f0`=P(K|NN), `f1`=P(K|ND) und `f2`=P(K|DD). Es wird davon ausgegangen, daß das Krankheitsallel D in einem Kopplungsungleichgewicht δ mit dem Markerallel 1 ist.

$$\texttt{delta} = P(D) - P(D) \cdot P(1)$$

Diese Gleichung läßt sich wie folgt umstellen.

$$P(D1) = \texttt{pm + delta}$$
$$P(D2) = \texttt{p(1-m) - delta}$$
$$P(N1) = \texttt{(1-p)m - delta}$$
$$P(N2) = \texttt{(1-p)(1-m) + delta}$$

Vorgegeben werden die Parameter `p`, m, `theta` und `delta`. Daraus werden durch Aufruf der Prozedur `PARAMET1` die Haplotypfrequenzen P(D1), P(N1), P(D2) und P(N2) berechnet.

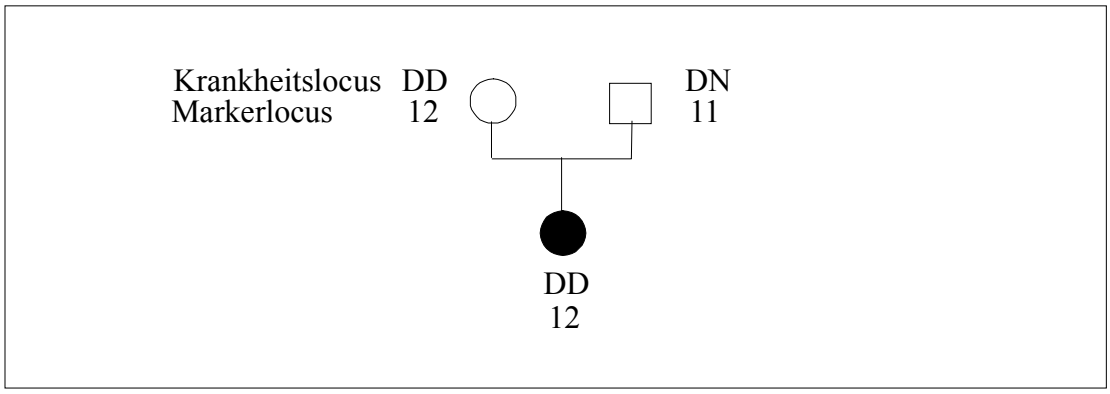
```
{d1, d2, n1, n2, tnt1} = PARAMET1(p, m, delta);
```

Die eigentliche Generierung eines Familientrios erfolgt dann durch Aufruf der Prozedur `FAMCREA1`.

```
{erg1} = FAMCREA1(d1, d2, n1, n2, theta, f0, f1, f2);
```

Eingabewerte sind die soeben berechneten Haplotypfrequenzen, die Rekombinationsrate `theta` und die Penetranzen `f0`, `f1` und `f2`. Ausgabewerte ist eine (2x8)-Matrix `erg1` mit den Genotypinformationen am Krankheits- und Markerlocus, die an folgendem Beispiel erklärt werden soll.

Abbildung 7-5 Beispiel für eine durch `FAMCREA1` generierte Familie mit bekannten Allelen am Krankheitslocus und Markerlocus



Die Genotyp- und die Transmissionsinformation werden jeweils getrennt für den Krankheitslocus und den Markerlocus in die Zeilen 1 und 2 der Matrix eingetragen. Die Markerallele 1 und 2 kodieren mit den entsprechenden Ziffern sowie das Allele D mit der Ziffern 1 und Allel N mit der Ziffer 2. In den Spalten 1 und 2 bzw. Spalten 3 und 4 werden. In den Spalten 5 und 6 bzw. 7 und 8 werden die Haplotypen der Eltern weiter in das transmittierte (T) und nicht transmittierte (NT) Allel differenziert. Dementsprechend ergibt sich für das Beispiel der folgende Eintrag in die Matrix.

Tabelle 7-2: (2x8)-Matrix mit den Genotypinformationen am Krankheits- und Markerlocus

	Vater		**Mutter**		**Paternaler Haplotyp**		**Maternaler Haplotyp**	
					T	NT	T	NT
Krankheitslocus	1	1	1	2	1	1	1	2
Markerlocus	1	2	1	1	2	1	1	1

Für das Erzeugen einer Familie werden zunächst vier auf dem Intervall [0;1] gleichverteilte Zufallszahlen z_1, …, z_4 gezogen. Gemäß der folgenden Entscheidungsregel werden die vier Haplotypen der Eltern festgelegt.

D1, falls $0 \leq z_i \leq P(D1)$

D2, falls $P(D1) < z_i \leq P(D1) + P(D2)$

N1, falls $P(D1) + P(D2) < z_i \leq P(D1) + P(D2) + P(N1)$

N2, falls $P(D1) + P(D2) + P(N1) < z_i \leq 1$

für i = 1,...,4.

Das Ergebnis wird in einer (2x4)-Matrix `mat` auf den Positionen [1,1] bis [2,4] abgespeichert. Die erste Zeile wird das Allel am Krankheitslocus, in der zweiten Zeile das Allel am Markerlocus eingetragen. Die Spalten 1 und 2 enthalten die Allele des Vaters und die Spalten 3 und 4 enthalten die Allele der Mutter.

Jetzt werden vier weitere Zufallszahlen z_5, …, z_8 benötigt zur Simulation von Transmission und Rekombination. Gemäß der folgenden Entscheidungsregel wird festgelegt, ob das Allel des Krankheitslocus an Position [1,1] oder an Position [1,2] an das Kind transmittiert wird:

Allel an Position [1,1], falls $z_5 \leq 0.5$

Allel an Position [1,2], falls $z_5 > 0.5$

Das transmittierte Allel wird in der Matrix an Position [1,5] eingetragen. Jetzt ist zu entscheiden, ob eine Rekombination stattgefunden hat oder nicht. ist $z_6 < (1-\theta)$, findet keine Rekombination statt, und das Markerallel in der gleichen Spalte wird transmittiert und an Position [2,5] eingetragen. Im anderen Fall wird das alternative Markerallel eingetragen. Die nicht transmittierten Allele am Krankheits- und Markerlocus werden

auf den Positionen [1,6] und [2,6] eingetragen. Analog wird bei der Transmission in der Mutter in den Spalten 7 und 8 verfahren.

Abhängig vom Genotyp am Krankheitslocus des Kindes wird danach unter Verwendung einer Zufallszahl z_9 entschieden, ob das Kind erkrankt oder gesund ist.

krank, falls $z_9 \leq f_i$.
gesund, falls $z_9 > f_i$
für i = 0,...2

Wird das Kind als krank klassifiziert, so wird die (2x8)-Matrix des Familientrios an die (2ix8)-Matrix `erg2` mit der Anzahl von i bisher erzeugten Familientrios angehängt. Die Anzahl der insgesamt simulierten Familien wird als Zähler mitgeführt. Zum Schluß erfolgt der Eintrag der Informationen am Markerlocus in die (2x2)-Matrix `tnt` (vgl. Tabelle 7-5). Ist das Kind gesund, wird die Familie aus dem Datensatz entfernt. Nach Generierung einer fixierten Anzahl Familien, die aus einem erkrankten Kind mit seinen beiden Eltern bestehen, wird die TDT-Teststatistik berechnet.

Nachteil des Ansatzes der zufälligen Generierung von Familientrios ist, daß die Programmschleife der Familiengenerierung so lange wiederholt werden muß, bis eine Familie mit einem erkrankten Kind simuliert wurde. Je seltener die Krankheit in einer Population auftritt, desto öfter muß im Durchschnitt die Programmschleife wiederholt werden.

7.3.2 Familiengenerierung mit `FAMCREA2`

Der zweite Ansatz ist in der Lage, Familientrios mit einem erkrankten Kind direkt unter Verwendung der Parameter `p`, `m`, `theta` und `delta` und der Penetranzen `f0`, `f1` und `f2` zu erzeugen. Dazu wird die Prozedur `PARAMET2` aufgerufen.

```
{psb, fam} = PARAMET2(p, m, delta, theta, f0, f1, f2);
```

Ausgegeben wird die (10x1)-Matrix `psb` mit bedingten Wahrscheinlichkeiten für den Genotyp am Markerlocus in Familien mit einem erkrankten Kind. Die (20x4)-Matrix `fam` enthält die jeweils passende Genotypkonstellation entsprechend der klassischen T/NT-Tabelle (vgl. Tabelle 6-2).

Für die direkte Generierung von Familientrios mit einem erkrankten Kind werden die Gleichungen aus Tabelle 3 in der Arbeit von KNAPP *et al.* (1993) verwendet. Tabelle 3

gibt die bedingten Wahrscheinlichkeiten der transmittierten und nicht transmittierten Markerallele eines erkrankten Kindes an. Dazu definierten KNAPP *et al.* (1993) zuerst die Variablen A und B und leiten daraus Tabelle 7-3 her.

$$A = \frac{p(f2-f1)+(1-p)(f1-f0)}{p^2 f2+2p(1-p)f1+(1-p)f0}\cdot delta^2$$

$$B = \frac{(f2-f1)-(f1-f0)}{p^2 f2+2p(1-p)f2+(1-p)f0}\cdot delta^2$$

Tabelle 7-3: Bedingte Wahrscheinlichkeiten für transmittierte und nicht transmittierte Markerallele in Familien mit einem erkrankten Kind (KNAPP *et al.*, 1993)

Nicht transmittierter Genotyp	**Transmittierter Genotyp**		
	11	12	22
11	m^4 $+2m^3A+m^2B$	$2m^3(1-m)$ $+2m^2(1-2m)A-2m^2B$ $+2\theta m^2A+2\theta mB$	$m^2(1-m)^2$ $-2m^2(1-m)A+m^2B$ $+2\theta m(1-m)A-$ $2\theta mB+\theta^2B$
12	$2m^3(1-m)$ $+4m^2A+2(1-m)B$ $-2\theta m^2A-2\theta mB$	$4m^2(1-m)^2$ $+4m(1-m)(1-2m)A$ $-4m(1-m)B+2\theta(1-\theta)B$	$2m(1-m)^3$ $-4m(1-m)^2A$ $+2m(1-m)B-$ $2\theta(1-m)^2A-2\theta(1-m)B$
22	$m^2(1-m)^2$ $+2m(1-m)^2A$ $+(1-m)^2B-2\theta m(1-m)A$ $-2\theta(1-m)B+\theta^2B$	$2m(1-m)^3$ $+2(1-m)^2(1-2m)A-$ $2(1-m)^2B-2\theta(1-m)^2A$ $+2\theta(1-m)B$	$(1-m)^4$ $+2(1-m)^3A+(1-m)^2B$

In die T/NT-Tabelle werden jedoch für ein erkranktes Kind die transmittierten und nicht transmittierten Allele getrennt in väterliche und mütterliche Haplotypen eingetragen. In Tabelle 7-3 wird aber eine solche Trennung nicht vorgenommen. Deshalb sind einige Erweiterungen notwendig. Die Werte für die einzelnen Einträge werden in der Prozedur `PARAMET2` berechnet und durch die Variablen `x1` - `x10` (vgl. Tabelle 7-4) bezeichnet.

Für folgende elterliche Genotypen ist die Zuordnung eindeutig, da beide Elternteile identische Genotypen besitzen müssen.

P(T 11 , NT 11 | Kind krank) ≙ x1

P(T 22 , NT 11 | Kind krank) ≙ x3

P(T 11 , NT 22 | Kind krank) ≙ x8

P(T 22 , NT 22 | Kind krank) ≙ x10

In folgenden Fällen muß ein Elternteil homozygot und ein Elternteil heterozygot sein, so daß ebenfalls eine eindeutige Zuordnung möglich ist. Die Einteilung in väterlichen und mütterlichen Genotyp erfolgt zufällig mit einer Chance von 1:1.

P(T 12 , NT 11 | Kind krank) ≙ x2

P(T 11 , NT 12 | Kind krank) ≙ x4

P(T 22 , NT 12 | Kind krank) ≙ x7

P(T 12 , NT 22 | Kind krank) ≙ x9

Für den Fall

P(T 12 , NT 12 | Kind krank)

ist die Zuordnung jedoch nicht eindeutig. In diesem Fall können die Eltern entweder beide heterozygot oder beide homozygot sein.

Sind beide Eltern homozygot, wird der Summand mit der Rekombinationsrate weggelassen und der verbleibende Term durch 2 geteilt.

P(T 12 , NT 12 | Eltern homozygot | Kind krank)

$= 2m^2(1-m)^2 + 2m(1-m)(1-2m)A - 2m(1-m)B \hat{=} x5$

Sind beide Eltern heterozygot, wird der Summand mit der Rekombinationsrate beibehalten.

P(T 12 , NT 12 | Eltern heterozygot | Kind krank)

$= 2m^2(1-m)^2 + 2m(1-m)(1-2m)A - 2m(1-m)B + 2\text{theta}(1-\text{theta})B \hat{=} x6$

Daraus ergibt sich folgende modifizierte Tabelle.

Tabelle 7-4 Modifizierte T/NT-Tabelle nach Knapp *et al.* (1993)

NT	**T**		
	11	12	22
11	x1	x2	x3
12	x4	x5 (Eltern homozygot)	x7
		x6 (Eltern heterozygot)	
22	x8	x9	x10

Die (20x4)-Matrix `fam` enthält die passenden Genotypen zu den Fällen `x1 - x10`. Die relativen Wahrscheinlichkeiten `x1 - x10` werden für die (1x10)-Matrix `psb` entsprechend folgender Regel addiert, so daß die Wahrscheinlichkeiten insgesamt auf dem Intervall [0;1] gleich verteilt sind.

```
psb [1,1]= x1
psb [1,2]= x1 + x2
psb [1,3]= x1 + x2 + x3
...
psb [1,10]= x1 + x2 + ...+ x10 = 1
```

Die Generierung der Familien erfolgt in der Prozedur `FAMCREA2` unter Verwendung der Matrizen `fam` und `psb`.

```
{erg3} = FAMCREA2(psb, fam);
```

Dazu wird eine Zufallszahl aus dem Intervall [0,1] gezogen und dem entsprechenden Intervall aus `psb` zugeordnet. Durch Zuordnung des zufällig gezogenen Intervalls zur Matrix `fam` enthält man die passenden Genotypen jeweils getrennt für Vater und Mutter. In der Matrix `fam` sind für jedes Intervall zwei Konstellationen der Genotypen vorgesehen, weil bei verschiedenen Genotypen der Eltern keine eindeutige Zuordnung zum Vater und zur Mutter möglich ist. Deshalb wird eine weitere Zufallszahl aus dem Intervall [0,1] gezogen, um verschiedene Genotypen mit einer Chance von 1:1 auf die beiden Eltern zu verteilen. Nach erfolgreicher Generierung der transmittierten und nicht transmittierten Genotypen des Kindes, getrennt nach paternaler und maternaler Herkunft, wird das Ergebnis in die T/NT-Tabelle eingetragen und an das Hauptprogramm übergeben.

Die Prozedur `TNTTAB` bestimmt die (1x4)-Matrix `tnt`, die der klassischen T/NT-Tabelle (vgl. Tabelle 6-2) entspricht, für `n` Familien aus einer (nx4)-Matrix `erg4`.

```
{tnt} = TNTTAB(erg4, n);
```

Tabelle 7-5: (1x4)-Matrix `tnt` der transmittierten und nicht transmittierten Markerallele für ein erkranktes Kind

Paternal		**Maternal**	
Transmittiert	Nicht transmittiert	Transmittiert	Nicht transmittiert
a	b	c	d

Die Zuordnung erfolgt für jede Familie i = 1,..., n nach den Gleichungen

```
tnt[erg4[i,1], erg4[i,2]] = tnt[erg4[i,1], erg4[i,2]]+1
tnt[erg4[i,3], erg4[i,4]] = tnt[erg4[i,3], erg4[i,4]]+1
```

7.3.3 Berechnung der TDT Statistik mit `Value`

Mit der Matrix `tnt` wird dann in der Prozedur `VALUE` die TDT-Statistik `tdtstat` berechnen.

```
{tdtstat} = VALUE(tnt);
tdtstat = ((tnt[1,2]-tnt[2,1]^2)/(tnt[1,2]+tnt[2,1]))
```

7.3.4 Partielles Löschen von Daten mit `Missings`

Um die Situation fehlender Daten zu simulieren, wird bei einigen Familien eine elterliche Genotypinformation gelöscht. Ohne Einschränkungen wird für die unvollständigen Trios angenommen, daß die Genotypen der Mutter bekannt und die des Vaters unbekannt sind. Dieses Vorgehen ist zulässig, weil die Generierung der Genotypen geschlechtsunabhängig erfolgt. Zur Simulation fehlender Genotypen wird die Prozedur `MISSINGS` verwendet.

Mit `MISSINGS` wird ein vorgegebener Anteil `v` an Vätern aus dem Datensatz von `n` vollständigen Familien `erg4` zufällig gelöscht. Der Aufruf der Prozedur erfolgt durch:

```
{ergm, pairs} = MISSINGS (erg4, n, v);
```

Auf dem Intervall [0;1] werden dazu n gleichverteilte Zufallszahlen z_1, …, z_v gezogen. Diesen Zufallszahlen werden in der Reihenfolge ihrer Ziehung die Reihe der ganzen Zahlen 1 bis `n` elementweise zugeordnet. Danach werden die beiden gekoppelten

Reihen nach der Größe der Zufallszahlen aufsteigend geordnet und die zugeordneten Nummern der ersten `v` Zufallszahlen geben die Zeilennummern der Familien an, in denen der Vater zu löschen ist. Aus diesen Familien wird eine (vx4)-Matrix `pairs` gebildet.

Die Reihen 1 und 2 erhalten die Informationen des Genotyps am Markerlocus des Kindes und die Reihen 3 und 4 erhalten die Informationen des Genotyps am Markerlocus der Mutter. Die Prozedur gibt die verbleibenden vollständigen Trios als Matrix `ergm` und die Matrix `pairs` mit den Informationen der unvollständigen Familien zurück.

Eine andere Möglichkeit wäre, die Genotypen der Väter entsprechend der Häufigkeiten zu löschen. Das heißt, daß ein häufiger Genotyp z.B. wahrscheinlicher gelöscht wird als ein seltener Genotyp. Sollten Genotypen allerdings systematisch fehlen, resultiert dieses in einer Verzerrung der Schätzungen.

7.4 Programmstruktur

7.4.1 Installation

Das Programm GAUSS© für WindowsNT/95 (Version 3.2.X) sollte vollständig auf der Festplatte installiert sein. Zur Installation des Programms müssen die Dateien der Programmdiskette in die entsprechenden Unterverzeichnisse auf die Festplatte kopiert werden. Die Dateien des im Rahmen dieser Arbeit entwickelten Computerprogramms sollten sich anschließend in folgenden Verzeichnisstrukturen befinden.

```
c:\\gauss\tdt\tdt.prg
c:\\gauss\tdt\crea2.src
c:\\gauss\tdt\eval2.src
c:\\gauss\tdt\rekonst2.src
```

Für die Ausgabe der Ergebnisse erfolgt in den Ordner.

```
c:\\gauss\tdt
```

Das entwickelte Programmsystem besteht aus dem Hauptprogramm `tdt.prg` und Quelldateien `crea2.src`, `eval2.src` und `rekonst2.src`. Im Hauptprogramm werden die Populationsparameter vorgegeben und verschiedene Prozeduren aus der

Quelldatei aufgerufen. Die Prozeduren werden unter Verwendung des `include`-Befehls aufgerufen.

7.4.2 Hauptprogramm

Das Hauptprogramm besteht aus folgenden Programmschritten

1. Format- und Ausgabebefehl für die Ergebnisse
2. Eingabe der Variablen
3. Berechnung bedingter Wahrscheinlichkeiten durch Aufruf von `PARAMET2`
4. Generierung von `n` Familientrios durch Aufruf von `FAMCREA2`
5. Berechnung der TDT-Teststatistik für komplette Familientrios
6. Löschen des Anteils `v` an Vätern durch Aufruf von `MISSINGS`
7. Schätzung der Allel- und Genotypfrequenzen durch Aufruf von `ESTIMm2` und `ESTIMg2`
8. Rekonstruktion nach der EM-Allel- und EM-Genotyprekonstruktion durch Aufruf von `REKONSTR1` und `REKONSTR4`
9. Berechnung der TDT-Teststatistik für die rekonstruierten Familientrios
10. Berechnung der Teststatistik des Combined Test durch Aufruf von `REKONSTR7`
11. Vergleich der Testwerte mit dem nominellen Signifikanzniveau einer χ^2-Verteilung mit 1 Freiheitsgrad
12. Ausgabe der prozentualen Fälle, in denen das nominelle Signifikanzniveau überschritten wurde.

Die globalen Variablen für die Eingabe der Populationsparameter werden definiert als:

`n`	Anzahl der insgesamt zu erzeugenden Familien
`r`	Anzahl der Replikationen bei der Generierung von `n` Familien
`am1`	Anteil von Population 1 an der Gesamtpopulation
`v1`	Anteil der zu löschenden Väter
`p1`	Frequenz des Allels D am Krankheitslocus in Population 1
`p2`	Frequenz des Allels D am Krankheitslocus in Population 2
`m1`	Frequenz des Allels 1 am Markerlocus in Population 1
`m2`	Frequenz des Allels 2 am Markerlocus in Population 2
`f0`	Penetranz der Krankheit beim Genotyp NN
`f1`	Penetranz der Krankheit beim Genotyp ND oder DN

`f2` Penetranz der Krankheit beim Genotyp DD

`delta1` Kopplungsungleichgewicht in Population 1

`delta2` Kopplungsungleichgewicht in Population 2

`theta` Rekombinationsrate in beiden Populationen

Folgende Variablen werden im Programm als globale Variablen verwendet:

`psb1` Bedingte Wahrscheinlichkeiten in Population1 (vgl. Kapitel 7.3.2)

`psb2` Bedingte Wahrscheinlichkeiten in Population2 (vgl. Kapitel 7.3.2)

`fam` Alle mögliche Genotypen von einem erkrankten Kind und einem Elternteil

`erg3` (1x4)-Matrix der aktuell erzeugten Familien mit FAMCREA2

`erg4` (ix4)-Matrix aller bereits erzeugten Familien mit FAMCREA2

`tnt` (2x2)-Matrix der T/NT-Tabelle für die aktuell erzeugte Familie

`tnt` (2x2)-Matrix der T/NT-Tabelle für alle erzeugten Familien mit Famcrea2

`tntm` (2x2)-Matrix der T/NT-Tabelle für vollständigen Familien nach Löschen der Väter

`tntr1` (2x2)-Matrix der T/NT-Tabelle nach EM-Allelrekonstruktion

`tntr4` (2x2)-Matrix der T/NT-Tabelle nach EM-Genotyprekonstruktion

`tnts1` (2x2)-Matrix der T/NT-Tabelle nach Kombination von tntm und tntr1

`tnts4` (2x2)-Matrix der T/NT-Tabelle nach Kombination von tntm und tntr4

`tdtstat` (rx10)-Matrix der TDT-Testwerte für komplette, inkomplette und rekonstruierte Familien

`sig` Anzahl der Fälle, in denen der Testwert tdtstat das nominale Signifikanzniveau einer χ^2-Verteilung überschritten hat

`sig2` Anzahl der Fälle, in denen der Testwert z des Combined Test das nominale Signifikanzniveau einer χ^2-Verteilung überschritten hat

7.5 Programmtests und Monte-Carlo Simulationen

Die folgenden Programmtests wurden auf einem Computer mit der Ausstattung Intel® Pentium® 120 MHz und 48 MB Arbeitsspeicher durchgeführt.

7.5.1 Generierung der Familien

Zur Generierung der Familien wurden die beiden Prozeduren `FAMCREA1` und `FAMCREA2` vorgestellt. Die folgenden Programmtest untersuchen die Gleichwertigkeit und die Geschwindigkeit beider Prozeduren bei der Generierung der Familien.

Folgende Parameter wurden vorgegeben

p=0,4; m=0,2; theta=0,1; delta=0,1; f0=0; f1=0; f2=0

Die Methode `FAMCREA1` benötigt für die Generierung von 2000 Familien mit 1000 Wiederholungen 4h 52min 24s.

Die Einträge in der T/NT-Tabelle zeigen die durchschnittlichen Abweichungen von den Erwartungswerten.

Tabelle 7-6 Durchschnittliche Abweichung vom Erwartungswert für `FAMCREA1` für 2000 Familien und 1000 Replikationen

Transmittiert	**Nicht transmittiert**	
	1	2
1	0.1370	-0.3810
2	0.3090	-0.0650

und die Standardabweichungen betragen

Tabelle 7-7 Standardabweichung für `FAMCREA1` für 2000 Familien und 1000 Replikationen

Transmittiert	**Nicht transmittiert**	
	1	2
1	17,9285	28,7265
2	21,5915	29,9166

Durchschnittlicher TDT-Testwert	340,6463
Durchschnittliche Abweichung vom Erwartungswert	0,2207
Standardabweichung	33,0646
Damit ergibt sich ein Bias von	0,06%.

Für die gleichen Parameter benötigt die Methode FAMCREA2 für die Generierung von 2000 Familien mit 1000 Wiederholungen 50min 9s.

Die Einträge in der T/NT-Tabelle zeigen die durchschnittlichen Abweichungen von den Erwartungswerten.

Tabelle 7-8 Durchschnittliche Abweichung vom Erwartungswert für FAMCREA2 für 2000 Familien und 1000 Replikationen

Transmittiert	**Nicht transmittiert**	
	1	2
1	0,0060	-0,5140
2	-0,0530	0,5610

und die Standardabweichungen betragen

Tabelle 7-9 Standardabweichung für FAMCREA2 für 2000 Familien und 1000 Replikationen

Transmittiert	**Nicht transmittiert**	
	1	2
1	18,5516	29,0351
2	21,3489	30,7849

Durchschnittlicher TDT-Testwert	340,9067
Durchschnittliche Abweichung vom Erwartungswert	0,4812
Standardabweichung	32,6420
Damit ergibt sich ein Bias von	0,14%.

Ein Vergleich der beiden Methoden zeigt, daß die Methode FAMCREA2 bedeutend schneller als die Methode FAMCREA1 ist. Die Genauigkeit der Verfahren ist sehr ähnlich. Aus dem Grund werden in weiteren Simulationen die Familien mittels FAMCREA2 generiert.

7.5.2 Schätzung der Genotyp- und Allelfrequenzen

Die Genotyp- und Allelfrequenzen können wie bereits beschrieben auf zwei unterschiedliche Arten berechnet werden. Bei der ersten Variante mit dem EM-Algorithmus werden zuerst die Frequenzen bei den vollständigen Trios bestimmt,

ohne die Frequenzen bei den Paaren zu berücksichtigen. Diese Frequenzen werden dann den fehlenden Elternteile zugrunde gelegt. Jetzt werden unter den vollständigen Trios und den vervollständigten Paare erneut die Frequenzen berechnet. Diese neuen Frequenzen werden wiederum den fehlenden Elternteilen zugrunde gelegt. Dieser Vorgang wird so lange wiederholt bis die Differenz ε zwischen einer neuen und der vorangegangenen Frequenz kleiner als 10^{-5} ist. Dieses Vorgehen wird in einem Programmtest mit dem Vorgehen verglichen, bei dem von Anfang an alle vorhandenen Genotypen der Eltern, also sowohl bei den vollständigen Eltern als auch bei den Paaren, bei der Schätzung der Allel- und Genotypfrequenzen berücksichtigt werden.

Die Vergleiche wurden unter einem rezessiven Modell und den Parametern p=0,4; m=0,3; θ=0,4 und δ=0,01 bei 70 Familientrios und 30 Paaren durchgeführt. Im ersten Durchgang der ersten Methode wurde eine Frequenz des Allels1 von em=0,30125 berechnet. Nach fünf Durchgängen war die Differenz $\varepsilon<10^{-5}$ und em=0,29706. Beim zweiten Vorgehen, bei dem alle Genotypinformationen in die Berechnung mit einbezogen wurden, wurde übereinstimmend mit dem Endergebnis der ersten Vorgehensweise eine Allelfrequenz von em=0,29706 berechnet. Dieser Vergleich ist 10.000 mal wiederholt worden. Das erste Vorgehen benötigt 9min 5s und das zweite Vorgehen benötigt 8min 51s.

Bei der Berechnung der Genotypfrequenzen müssen 11 Einzelwahrscheinlichkeiten berechnet werden (vgl. Kapitel 7.2.5). Das erste Vorgehen wird so oft wiederholt bis die Differenzen ε aller Einzelwahrscheinlichkeiten kleiner 10^{-5} sind. Das zweite Vorgehen liefert auch hier die gleichen Ergebnisse. Für die gleiche Situation wie bei der Schätzung der Allelfrequenz em sind sechs Durchgänge notwendig. Ein Geschwindigkeitsvergleich zeigt bei 10.000 Wiederholungen, daß für das erste Verfahren 19min 15s und dem zweite Verfahren nur 18min 43s benötigt werden.

Bei gleichwertiger Frequenzschätzung wird aufgrund des Geschwindigkeitsvorteil das zweite Verfahren für die Frequenzschätzung verwendet.

7.5.3 Monte-Carlo Simulationen

Monte-Carlo (MC) Simulationen wurden auf einem PC mit der Ausstattung Intel® PentiumPro® Prozessor 200 MHz und 64 MB Arbeitsspeicher durchgeführt. Mehrere MC-Simulationssätze wurden bei 8 verschiedenen Graden von Populationsstratifikationen durchgeführt. Innerhalb eines MC-Simulationssatzes wurden

vier Vererbungsmodelle und drei verschiedene Rekombinationsraten berücksichtigt. Aus den vorgegebenen Allelfrequenzen p und m ergaben sich je 12 Kombinationsmöglichkeiten. Bei jeder Kombinationsmöglichkeit wurden bei drei verschiedene Anteile fehlender Genotypen die EM-Allel- und EM-Genotyprekonstruktionen durchgeführt. Im Durchschnitt nahm eine dieser MC-Simulationen 2 Tage und 14 Stunden Zeit in Anspruch. Insgesamt wurden 18 dieser MC-Simulationssätze berechnet. Für die MC-Simulationen mit dem 1-TDT wurden nur drei Kombinationsmöglichkeiten berücksichtigt, aber in einem MC-Simulationssatz alle Grade an Populationsstratifikation simuliert. Diese MC-Simulationen wurden auf einem PC mit der Ausstattung Intel© Pentium© III 600 MHz und 128 MB Arbeitsspeicher durchgeführt. Die MC-Simulationszeit betrug hierbei 10 Stunden und 46 Minuten. Die Ergebnisse der MC-Simulationen werden im folgenden für die einzelnen Vererbungsmodelle dargestellt.

8 Anwendung des Transmission-Disequilibrium Tests bei unvollständigen Daten

8.1 Monte-Carlo Simulationsmodelle

Die Güte der beiden EM-Rekonstruktionsmethoden wird durch Monte-Carlo (MC) Simulationen untersucht. Mit dem von mir geschriebenen Computerprogramm werden die Power unter der Alternativhypothese und das empirische Signifikanzniveau unter der Nullhypothese für beide EM-Rekonstruktionsverfahren geschätzt und mit dem 1-TDT verglichen. Die MC-Simulationen werden bei unterschiedlichen Graden an Populationsstratifikation, verschiedenen Vererbungsmodellen und Populationsparametern getestet. Die asymptotische Version der TDT-Teststatistik wird für die kompletten Familien, die Familien mit teilweise fehlenden Genotypen (vollständige Familien) und für die Familien nach Rekonstruktion der fehlenden Genotypen durch die beiden EM-Rekonstruktionsmethoden berechnet. In zusätzlichen MC-Simulationen wird das Programm für den 1-TDT erweitert. Für jede Konstellation der Parameter werden 100 Familien unter Verwendung der direkten Generierung mit FAMCREA2 ausgeführt. Danach erfolgt das zufällige Löschen der kompletten Genotypinformationen von 10%, 30% und 50% der Väter. Anschließend werden die fehlenden Daten mit der EM-Allel- der und EM-Genotyprekonstruktionsmethode rekonstruiert. Für eine hinreichend genaue Schätzung bei der Power- als auch der Signifikanzanalyse werden für jeden MC-Simulationsfall 10.000 Replikationen durchgeführt. Nach jeder Replikation wird überprüft, ob die berechnete TDT-Statistik den Wert von 3,8414588 entsprechend dem nominellen Signifikanzniveau von 5% für eine χ^2-Verteilung überschreitet. Der Prozentanteil der Replikationen, in denen dieser Wert überschritten wird, gibt unter den Bedingungen der Alternativhypothese die Power und unter den Bedingungen der Nullhypothese das empirische Signifikanzniveau an.

Die folgenden Parameter werden zur Generierung von Familien mit zwei Elternteilen und einem erkrankten Kind mit Kenntnis aller Genotypen am Markerlocus vorgegeben.

8.1.1 Vererbungsmodelle

Es werden vier Vererbungsmodelle mit den Penetranzen f_0, f_1, f_2 berücksichtigt. Die Penetranz f_0 bezieht sich auf den homozygoten Genotyp bei nicht mutiertem

Markerlocus. Die Penetranz f_1 bezieht sich auf einen heterozygoten Markerlocus und die Penetranz f_2 auf einen homozygoten Markerlocus.

Tabelle 8-1 Penetranzen der Vererbungsmodelle für die Monte Carlo-Simulationen

Vererbungsmodelle	**Penetranzen**		
	f_0	f_1	f_2
Rezessiv	0	0	1
Dominant	0	1	1
Additiv	0	0,5	1
Additiv mit Phänokopie	0,1	0,5	0,5

8.1.2 Bedingungen für Populationsstratifikation

Zur Berücksichtigung des Phänomens Populationsstratifikation wird die Gesamtheit der erzeugten Familien aus zwei Populationen gebildet. Die Populationen unterscheiden sich in der Assoziation zwischen Krankheitslocus und Markerlocus. Die verwendeten Indizes einzelner Parameter bezeichnen die Zugehörigkeit des Wertes zu den Population 1 oder 2. Wird kein Index verwendet, so gilt der Parameter für beide Populationen.

Familien der Population 1 weisen unter den Bedingungen der Alternativhypothese eine positive Assoziation und Kopplung auf, während Familien der Population 2 keine positive Assoziation zeigen. Damit ist für Population 2 immer die Bedingung der Nullhypothese erfüllt. Die Population 2 führt also zu Populationsstratifikation in der Gesamtpopulation. Der Anteil der Familien aus Population 1 beträgt in getrennten MC-Simulationen unter gleichbleibenden Bedingungen 0, 10, 20, 30, 50, 70, 90 und 100%. Bei positiver Assoziation in der Population 1 wird $\delta_1 = \delta_{max}$ gesetzt. Dabei bezeichnet δ_{max} die maximale Assoziation. Sie läßt sich aus den vorgegebenen Allelfrequenzen am Krankheitslocus und Markerlocus berechnen: Ist p größer oder gleich m, ist $\delta_{max} = m \cdot (1-p)$. Ist m größer als p, gilt $\delta_{max} = p \cdot (1-m)$. Die definierten Rekombinationsraten θ gelten für beide Populationen 1 und 2. Für die Allelfrequenzen p und m werden zwei Varianten vorgegeben.

8.1.3 Variante I der Populationsparameter

In Variante I sind die Allelfrequenzen p und m für beide Populationen gleich. Beide unterscheiden sich nur im Assoziationsparameter δ voneinander. Bei den MC-

Simulationen in Variante I gelten alle übrigen Parameter für beide Populationen gleichermaßen. Folgende Variationen der Parameter werden betrachtet:

Tabelle 8-2 **Variante I der Allelfrequenzen**
Population 1 und Population 2 unterscheiden sich nur in der Assoziation zwischen Markerlocus und Krankheitslocus

Population 1							**Population 2**					
Assoziationsparameter δ_1							Assoziationsparameter δ_2					
δ_{max}							0					
Rekombinationsrate θ												
0			0,01			0,5						
p	0,01			0,05			0,1			0,3		
m	0,1	0,3	0,5	0,1	0,3	0,5	0,1	0,3	0,5	0,1	0,3	0,5

Die Bedingungen der Nullhypothese sind für die Rekombinationsrate $\theta = 0{,}5$ erfüllt.

8.1.4 Variante II der Populationsparameter

In Variante II unterscheiden sich die beiden Populationen in den unterschiedlichen Markerallelfrequenzen m_1 und m_2. Die Krankheitsallelfrequenz p gilt für beide Populationen. Da die Markerallelfrequenzen in beiden Populationen unterschiedlich sind ergeben sich in bezug auf den Assoziationsparameter δ und die Rekombinationsrate θ mehrere Möglichkeiten für Monte–Carlo Simulationen unter der Nullhypothese und Alternativhypothese.

Insgesamt gibt es dann drei verschiedene Möglichkeiten für die Gültigkeit der Nullhypothese.

a) $\theta = 0{,}5$ und $\delta_1 \neq 0$

b) $\delta_1 = \delta_2 = 0$ und $\theta = 0{,}5$

c) $\delta_1 = \delta_2 = 0$ und $\theta \neq 0{,}5$

MC-Simulationen unter Nullhypothese für Möglichkeit a) werden wie bereits beschrieben in Variante I durchgeführt. In Variante IIa sind die Parameterkonstellationen für die gleichen Assoziationsparameter und die Rekombinationsraten wie in Variante I dargestellt. Auch hier sind die Bedingungen der Nullhypothese entsprechend Möglichkeit a) erfüllt.

Tabelle 8-3 Variante IIa: In Population 1 liegt Assoziation zwischen Markerlocus und Krankheitslocus vor

Population 1									**Population 2**							
Assoziationsparameter δ_1									Assoziationsparameter δ_2							
δ_{max}									0							
Rekombinationsrate θ																
0				0,01					0,5							
p	0,01				0,05				0,1				0,3			
m_1	0,1	0,1	0,3	0,3	0,1	0,1	0,3	0,3	0,1	0,1	0,3	0,3	0,1	0,1	0,3	0,3
m_2	0,2	0,3	0,4	0,5	0,2	0,3	0,4	0,5	0,2	0,3	0,4	0,5	0,2	0,3	0,4	0,5

Da sich hier im Gegensatz zu Variante I beide Populationen bereits in der Markerallelfrequenz unterscheiden und damit schon zu Populationsstratifikation führen, können in auch die Möglichkeit b) und c) berücksichtigt werden. In Variante IIb sind die Assoziationsparameter δ_1 und δ_2 gleich null.

Tabelle 8-4 Variante IIb: In beiden Populationen liegt keine Assoziation zwischen Markerlocus und Krankheitslocus vor

Population 1									**Population 2**							
Assoziationsparameter δ_1									Assoziationsparameter δ_2							
0									0							
Rekombinationsrate θ																
0				0,01					0,5							
p	0,01				0,05				0,1				0,3			
m_1	0,1	0,1	0,3	0,3	0,1	0,1	0,3	0,3	0,1	0,1	0,3	0,3	0,1	0,1	0,3	0,3
m_2	0,2	0,3	0,4	0,5	0,2	0,3	0,4	0,5	0,2	0,3	0,4	0,5	0,2	0,3	0,4	0,5

8.1.5 Variante III der Populationsparameter

Weitere MC-Simulationen werden zum Vergleich der EM-Rekonstruktionsmethoden mit dem 1-TDT (SUN *et al.*, 1999) durchgeführt. Dazu werden in Variante III drei Parameterkonstellationen der Variante I erneut betrachtet.

Tabelle 8-5 Variante III: Vergleich der EM-Rekonstruktionsmethoden mit dem 1-TDT, Population 1 und Population 2 unterscheiden sich nur in der Assoziation zwischen Markerlocus und Krankheitslocus

Population 1			**Population 2**
Assoziationsparameter δ_1			Assoziationsparameter δ_1
δ_{max}			0
Rekombinationsrate θ			
0			0,5
p	0,01	0,05	0,3
m	0,1	0,3	0,1

8.2 Ergebnisse unter der Nullhypothese

MC-Simulationen wurden unter der Nullhypothese bei Fehlen von Assoziation und/oder Kopplung in der Variante I für θ=0,5; $\delta_1=\delta_{max}$; δ_2=0 und in Variante II für die Möglichkeiten a) θ=0,5; $\delta_1=\delta_{max}$; δ_2=0, b) θ=0,5; $\delta_1=\delta_2$=0 und c) θ=0; $\delta_1=\delta_2$=0 bzw. θ=0,01; $\delta_1=\delta_2$=0.

8.2.1 Rezessives Vererbungsmodell

Für das rezessive Vererbungsmodell wurden die Penetranzen f_0=0; f_1=0; f_2=1 festgelegt. Die Ergebnisse der empirischen Signifikanzniveaus der MC-Simulationen für Variante I, II und III durchgeführt. Im ersten Schritt werden die empirischen Signifikanzniveaus bei der Berechnung der 1-TDT Statistik und der klassischen TDT Statistik für alle Familien, vollständige Familien, die EM-Allelrekonstruktion und die EM-Genotyprekonstruktion vergleichen.

In Abbildung 8-1 sind die Ergebnisse aus Variante I für die Populationsparameter p=0,3; m=0,1; δ_1=0,07; δ_2=0 und θ=0,5 dargestellt. Der Anteil fehlender Genotypinformationen der Väter beträgt 10%.

Abbildung 8-1 **Empirische Signifikanzniveaus (in Prozent)**
Transmission-Disequilibrium Tests mit allen Familien, vollständigen Familien, EM-Allelrekonstruktion, EM-Genotyprekonstruktion und 1-TDT bei einem nominellen Signifikanzniveau von 5%
Globale Modellparameter: p=0,3; m=0,1; δ_1=0,07; δ_2=0; θ=0,5
Rezessives Vererbungsmodell: f_0=0; f_1=0; f_2=1

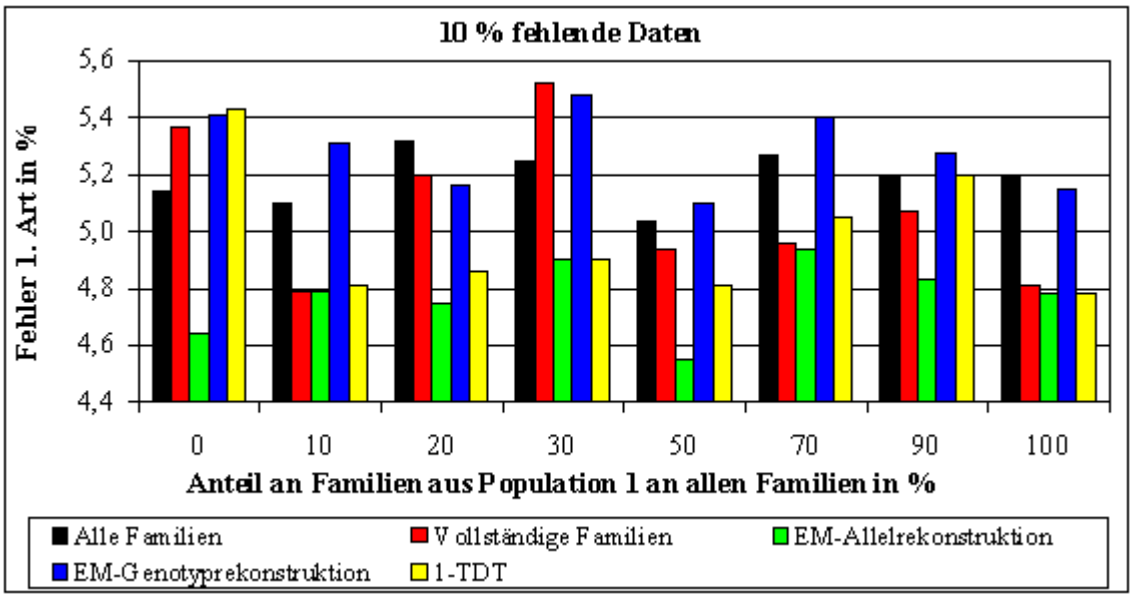
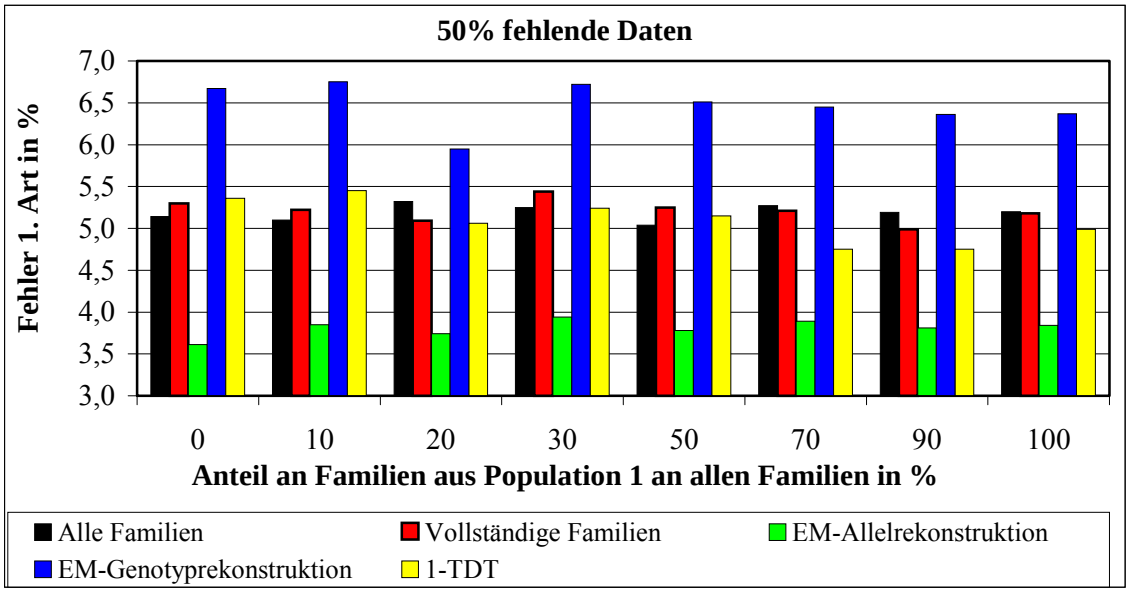


Abbildung 8-1 zeigt, daß bei einem geringen Anteil fehlender Genotypen das empirische Signifikanzniveau aller Tests im Intervall von 4,5% bis 5,5% liegt. Die EM-Allelrekonstruktion unterbietet das nominelle Niveau von 5% bei allen Graden an Populationsstratifikation. Das empirische Niveau der EM-Genotyprekonstruktion dagegen liegt in allen Fällen über 5%, mit einem Maximalwert von 5,48%. Ähnlich verhält sich das empirische Niveau für den TDT bei allen Familien, mit einem Maximalwert von 5,32%. Das empirische Niveau des 1-TDT variiert im Intervall von 4,78% bis 5,43%, das empirische Niveau des TDT für vollständige Familien im Intervall von 4,79% bis 5,52%.

In Abbildung 8-2 ist das Ergebnis bei gleicher Parameterkonstellation für einen mit 50% höheren Anteil fehlender Genotypen dargestellt.

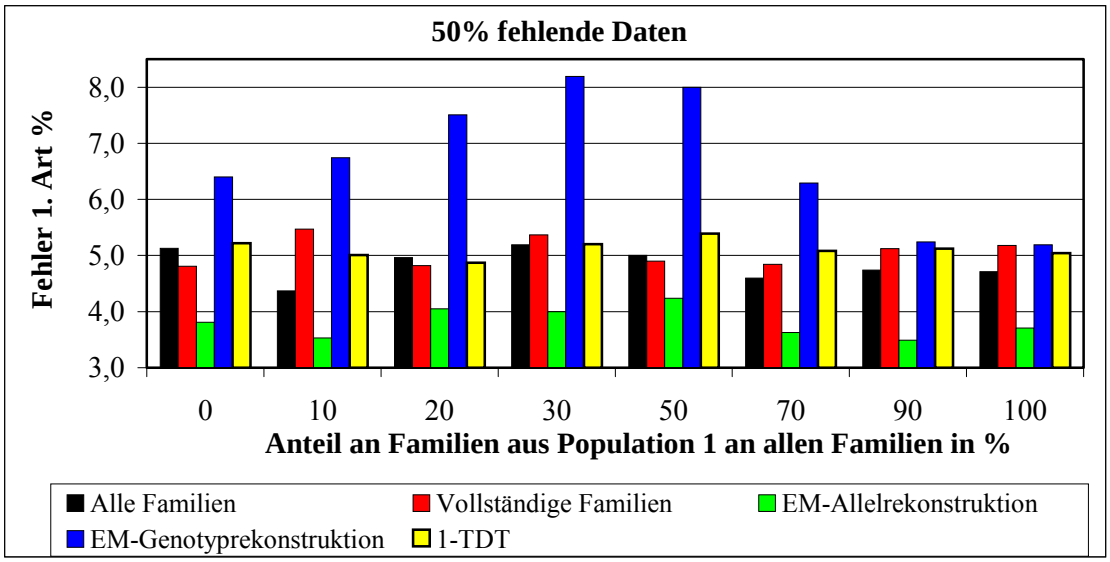
Abbildung 8-2 **Empirische Signifikanzniveaus (in Prozent)**
Transmission-Disequilibrium Test mit allen Familien, vollständigen Familien, EM-Allelrekonstruktion, EM-Genotyprekonstruktion und 1-TDT bei einem nominellen Signifikanzniveau von 5%
Globale Modellparameter: $p=0,3$; $m=0,1$; $\delta_1=0,07$; $\delta_2=0$; $\theta=0,5$
Rezessives Vererbungsmodell: $f_0=0$; $f_1=0$; $f_2=1$



Verglichen mit dem Ergebnis aus Abbildung 8-1 halten der TDT für vollständige Familien und der 1-TDT in etwa das nominelle Niveau von 5%. Die empirischen Signifikanzwerte liegen bei den vollständigen Familien im Intervall von 4,99% bis 5,44% und beim 1-TDT im Intervall von 4,75% bis 5,45%. Das empirische Signifikanzniveau der EM-Genotyprekonstruktion dagegen überschreitet für jeden Grad der Populationsstratifikation das nominelle Niveau von 5% deutlich. Die empirischen Signifikanzwerte liegen im Intervall von 5,95% bis 6,75%. Damit hält die EM-Genotyprekonstruktion das nominelle Niveau nicht und ist zu liberal. Die empirischen Signifikanzen der EM-Allelrekonstruktion liegen im Intervall von 3,61% bis 3,94% und tendieren damit zur Konservativität.

Zum Vergleich zeigt Abbildung 8-3 ebenfalls für 50% fehlende Genotypen die empirischen Signifikanzen für die Populationsparameter p=0,01; m=0,1; δ_1=0,009; δ_2=0 und θ=0,5.

Abbildung 8-3 **Empirische Signifikanzniveaus (in Prozent)**
Transmission-Disequilibrium Test mit allen Familien, vollständigen Familien, EM-Allelrekonstruktion, EM-Genotyprekonstruktion und 1-TDT bei einem nominellen Signifikanzniveau von 5%
Globale Modellparameter: p=0,01; m=0,1; δ_1=0,009; δ_2=0; θ=0,5
Rezessives Vererbungsmodell: f_0=0; f_1=0; f_2=1

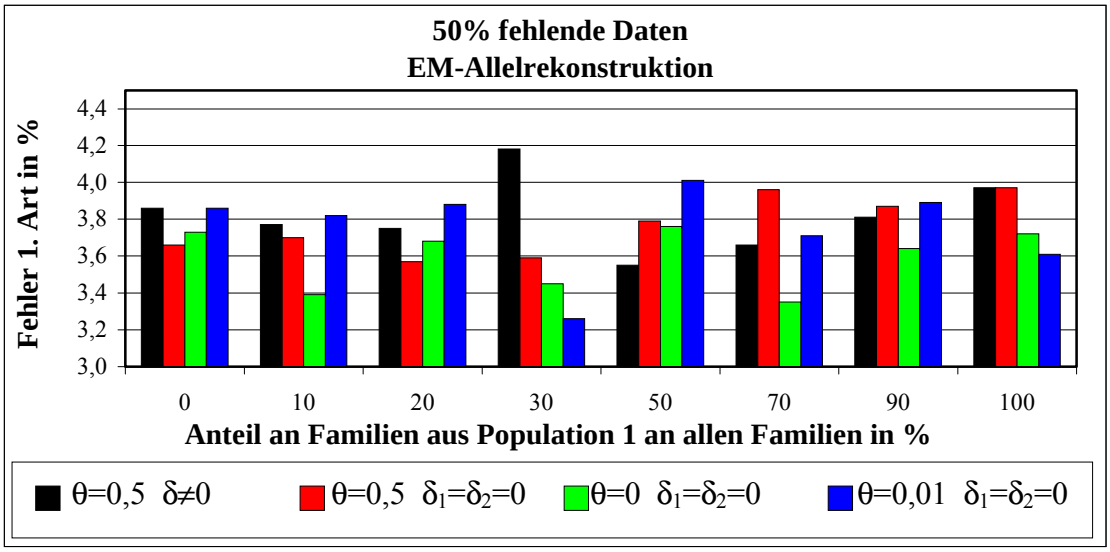


Analog zu Abbildung 8-2, halten der TDT für vollständige Familien und der 1-TDT das nominelle Niveau von 5% für das in Abbildung 8-3 dargestellte Modell. Die EM-Genotyprekonstruktion hält zwar in diesem Fall das nominelle Niveau bei einem Grad von 90% und 100% an Populationsstratifikation, aber bei 30% Populationsstratifikation erreicht das nominelle Signifikanzniveau einen Maximalwert von 8,19%. Das heißt, die EM-Genotyprekonstruktion ist auch hier in einigen Fällen zu liberal. Bei der EM-Allelrekonstruktion werden vergleichbare empirische Signifikanzen beobachtet wie in Abbildung 8-2. Die empirischen Signifikanzwerte liegen im Intervall von 3,49% bis 4,00%. Die EM-Allelrekonstruktion ist damit auch in diesem Fall zu konservativ.

Für die EM-Genotyp- und EM-Allelrekonstruktion zeigen Abbildung 8-4 und Abbildung 8-5 die Fehler 1. Art für die Variante IIb. In Variante IIb wurden für beide Populationen unterschiedliche Markerallelfrequenzen vorgegeben. Unter Beibehaltung von Populationsstratifikation waren daher MC-Simulationen unter drei unterschiedlichen Konstellationen der Nullhypothese möglich. Als feste Populationsparameter wurden vorgegeben p=0,01; m_1=0,1; m_2=0,2 und δ_2=0.

Abbildung 8-4 zeigt für 50% fehlende Genotypen die empirischen Signifikanzen bei der TDT mit EM-Allelrekonstruktion unter vier Bedingungen für die Nullhypothese.

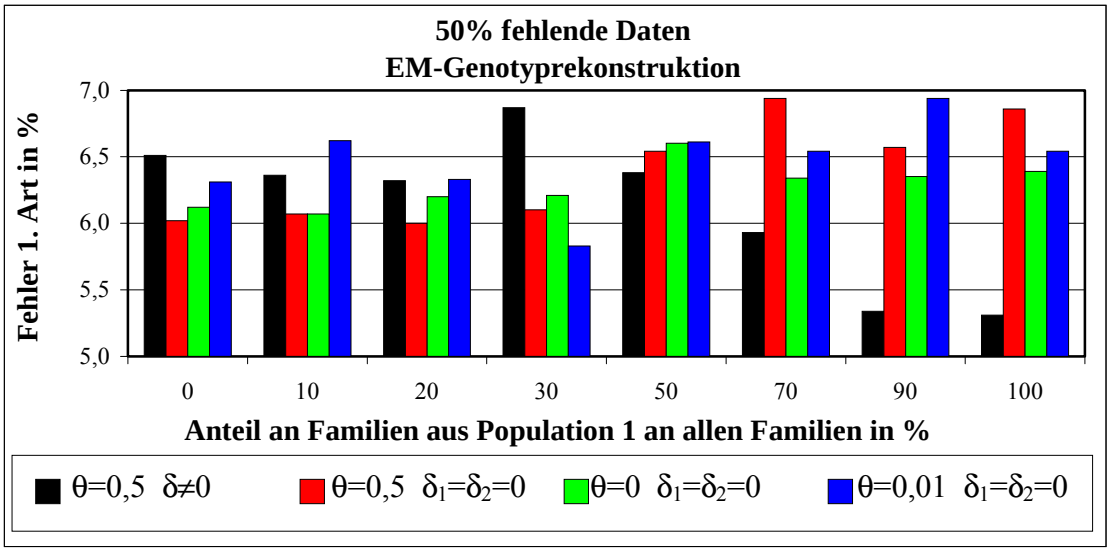
Abbildung 8-4 **Empirische Signifikanzniveaus (in Prozent)**
Transmission-Disequilibrium Test mit EM-Allelrekonstruktion
unter verschiedenen Bedingungen der Nullhypothese H_0
bei einem nominellen Signifikanzniveau von 5%
Globale Modellparameter: p=0,01; m=0,1; m2=2; δ_2=0
Rezessives Vererbungsmodell: f_0=0; f_1=0; f_2=1



Bei der EM-Allelrekonstruktion zeigt Abbildung 8-4, daß in keinem Fall das nominelle Signifikanzniveau von 5% überschritten wird. Alle empirischen Signifikanzen liegen im Intervall von 3,26% bis 4,18%.

Abbildung 8-5 zeigt für 50% fehlende Genotypen die empirischen Signifikanzen bei der TDT mit EM-Genotyprekonstruktion unter vier Bedingungen für die Nullhypothese.

Abbildung 8-5 **Empirische Signifikanzniveaus (in Prozent)**
Transmission-Disequilibrium Test mit EM-Genotyprekonstruktion
unter verschiedenen Bedingungen der Nullhypothese H_0
bei einem nominellen Signifikanzniveau von 5%
Globale Modellparameter: p=0,01; m1=0,1; m2=0,2; δ_2=0
Rezessives Vererbungsmodell: f_0=0; f_1=0; f_2=1

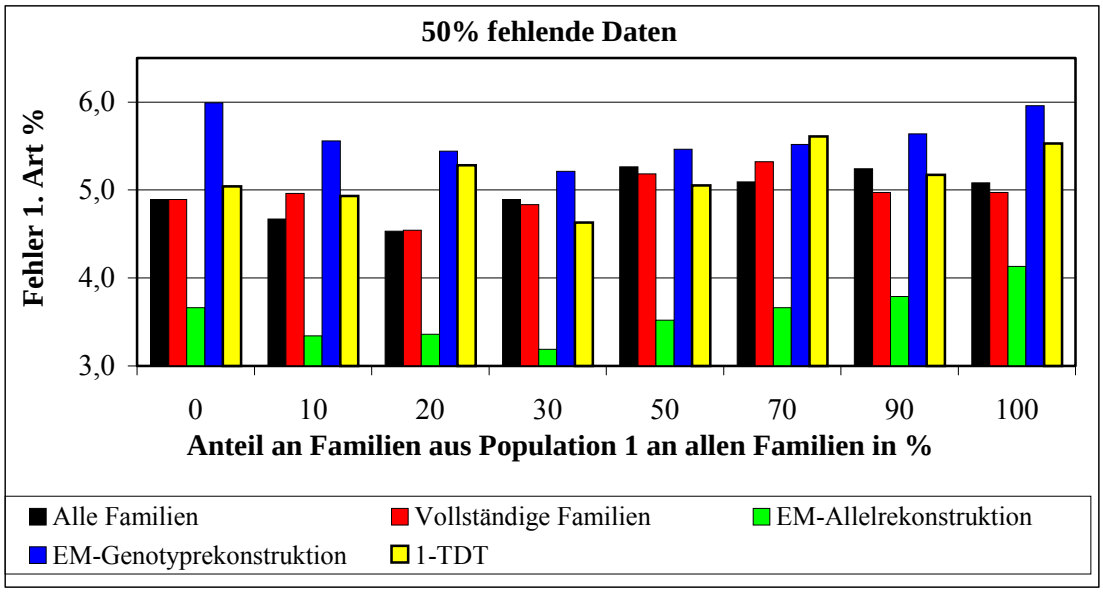


Dagegen überbieten die empirischen Niveaus bei der EM-Genotyprekonstruktion in allen Fälle das nominelle Signifikanzniveau von 5%. Alle empirischen Signifikanzen liegen im Intervall von 6,94% bis 5,31%.

8.2.2 Dominantes Vererbungsmodell

Für das dominante Vererbungsmodell mit den Penetranzen f_0=0; f_1=1; f_2=1 werden zuerst die empirischen Signifikanzniveaus bei der Berechnung der 1-TDT Statistik und der klassischen TDT Statistik für alle Familien, vollständige Familien, die EM-Allelrekonstruktion und die EM-Genotyprekonstruktion verglichen. Abbildung 8-6 zeigt die Ergebnisse aus Variante I für die Populationsparameter p=0,05; m=0,3; δ_1=0,035; δ_2=0 und θ=0,5 dargestellt. Der Anteil fehlender Genotypinformationen der Väter beträgt 50%.

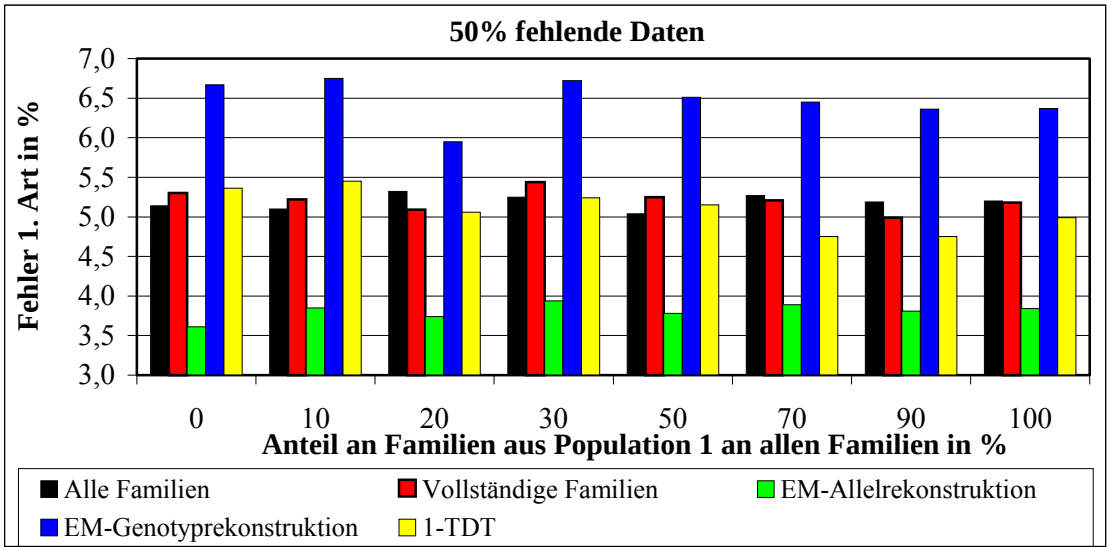
Abbildung 8-6 **Empirische Signifikanzniveaus (in Prozent)**
Transmission-Disequilibrium Test mit allen Familien, vollständigen Familien, EM-Allelrekonstruktion, EM-Genotyprekonstruktion und 1-TDT bei einem nominellen Signifikanzniveau von 5%
Globale Modellparameter: p=0,05 m=0,3 δ_1=0,035 δ_2=0 θ=0,5
Dominantes Vererbungsmodell: f_0=0; f_1=1; f_2=1



Bei einem Anteil von 50% fehlender Genotypen streut das empirische Signifikanzniveau die empirische Signifikanz bei den kompletten Familien im Intervall von 4,53% bis 5,26%. Die EM-Allelrekonstruktion unterbietet das nominelle Niveau von 5% bei allen Graden von Populationsstratifikation. Die empirischen Signifikanzniveaus liegen im Intervall von 3,19% bis 4,13%. Das Niveau der EM-Genotyprekonstruktion dagegen liegt in allen Fällen über 5%, mit einem Maximalwert von 5,99%. Das empirische Niveau für den TDT bei vollständigen Familien streut im Intervall von 4,45% bis 5,32%, das empirische Niveau des 1-TDT im Intervall von 4,63% bis 5,61%.

In Abbildung 8-7 ist das Ergebnis ebenfalls für 50% fehlende Genotypen aber bei der Parameterkonstellation p=0,3; m=0,1; δ_1=0,07; δ_2=0 und θ=0,5 dargestellt.

Abbildung 8-7 **Empirische Signifikanzniveaus (in Prozent)**
Transmission-Disequilibrium Test mit allen Familien, vollständigen Familien, EM-Allelrekonstruktion, EM- Genotyprekonstruktion und 1-TDT bei einem nominellen Signifikanzniveau von 5%
Globale Modellparameter: p=0,3 m=0,1 δ_1=0,07 δ_2=0 θ=0,5
Dominantes Vererbungsmodell: f_0=0; f_1=1; f_2=1

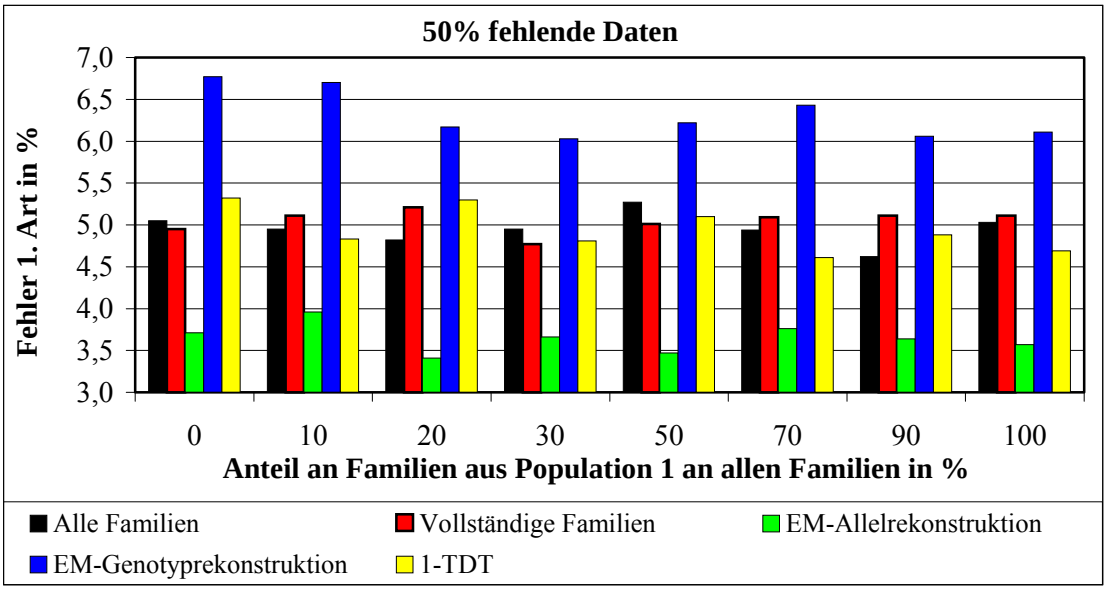


Ähnlich wie in Abbildung 8-6 zu sehen, halten der TDT für vollständige Familien und der 1-TDT das nominelle Niveau von 5%. Die EM-Genotyprekonstruktion überbietet in diesem Fall das nominelle Niveau sehr deutlich und zeigt einen Maximalwert von 6,75%. Die EM-Genotyprekonstruktion kann in beiden Fällen bei allen Graden von Populationsstratifikation das nominelle Niveau nicht halten und ist zu liberal. Bei der EM-Allelrekonstruktion wird ein vergleichbares empirische Signifikanzniveau beobachtet wie in Abbildung 8-6. Die empirischen Signifikanzwerte liegen im Intervall von 3,61% bis 3,94%. Die EM-Allelrekonstruktion ist damit auch bei der dominanten Vererbung Fall zu konservativ.

8.2.3 Additives Vererbungsmodell

In Abbildung 8-8 sind die empirischen Signifikanzen für ein additives Vererbungsmodell mit den Penetranzen betragen f_0=0, f_1=0,5 und f_2=1 und den Parameter p=0,3; m=0,1; δ_1=0,07; δ_2=0 und θ=0,5 bei 50% fehlenden Daten dargestellt.

Abbildung 8-8 **Empirische Signifikanzniveaus (in Prozent)**
Transmission-Disequilibrium Test mit allen Familien, vollständigen Familien, EM-Allelrekonstruktion, EM-Genotyprekonstruktion und 1-TDT bei einem nominellen Signifikanzniveau von 5%
Globale Modellparameter: p=0,3 m=0,1 δ_1=0,07 δ_2=0 θ=0,5
Additives Vererbungsmodell: f_0=0; f_1=0,5; f_2=1



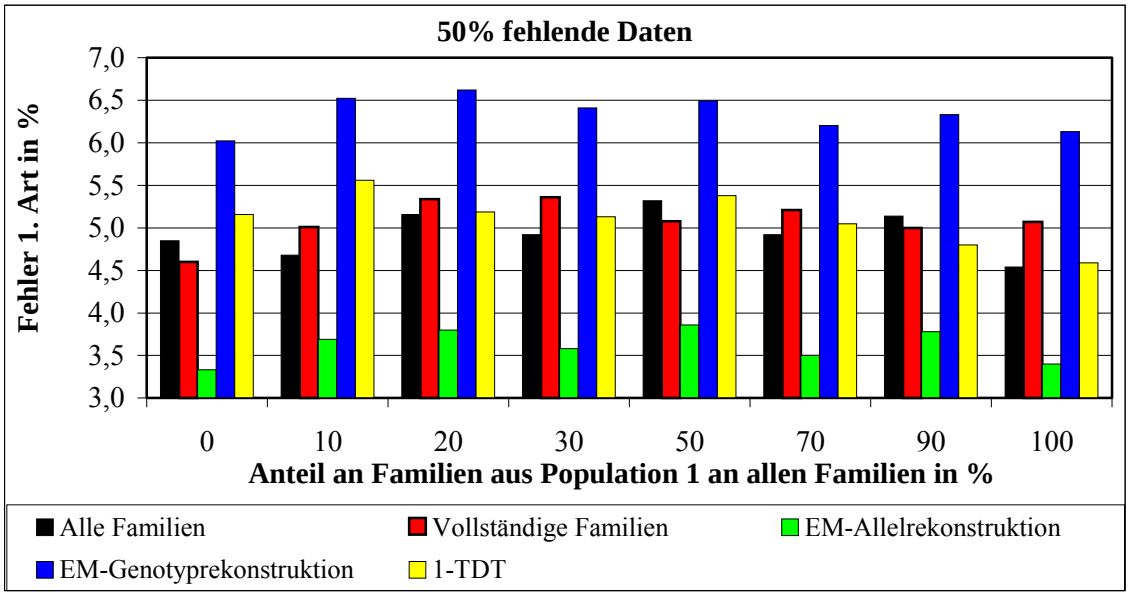
Vergleichbar mit den bisherigen Ergebnissen hält auch in diesem Fall die EM-Genotyprekonstruktion bei allen Graden an Populationsstratifikation das nominelle Niveau von 5% nicht. Die empirischen Signifikanzen variieren im Intervall von 6,03% bis 6,77%. Bei der EM-Allelrekonstruktion sind die empirischen Signifikanzen dagegen auch hier deutlich unterhalb der Schwelle von 5% und streuen im Intervall von 3,41 bis 3,96. Die empirischen Signifikanzniveaus des TDT für vollständige Familien und des 1-TDT korrelieren mit dem nominellen Niveau von 5%.

8.2.4 Additives Vererbungsmodell mit Phänokopie

Die letzten MC-Simulationen wurden unter einem additiv dominanten Modell mit den Penetranzen f_0=0,1; f_1=0,5 und f_2=0,5 durchgeführt. Die Phänokopierate entspricht der Penetranz f_0=0,1. Daß heißt 10% der Individuen ohne das vermeintliche Krankheitsgen entwickeln dennoch den gleichen Phänotyp wie bei der Krankheit. Auch hier werden

die Ergebnisse für die Parameter p=0,3; m=0,1; δ_1=0,07; δ_2=0 und θ=0,5 bei 50% fehlenden Genotypinformationen vorgestellt. Die erste Abbildung stellt die verschiedenen empirischen Signifikanzniveaus dar.

Abbildung 8-9 **Empirische Signifikanzniveaus (in Prozent)**
Transmission-Disequilibrium Test mit allen Familien, vollständigen Familien, EM-Allelrekonstruktion, EM- Genotyprekonstruktion und 1-TDT bei einem nominellen Signifikanzniveau von 5%
Globale Modellparameter: p=0,3 m=0,1 δ_1=0,07 δ_2=0 θ=0,5
Additives Vererbungsmodell mit Phänokopie: f_0=0,1; f_1=0,5; f_2=0,5



Auch bei diesem Vererbungsmodell ergeben sich ähnliche empirische Signifikanzniveaus wie unter den bisher dargestellten Vererbungsmodellen. Das empirischen Niveau eines Fehlers 1. Art liegt für die EM-Genotyprekonstruktion im Intervall von 6,02% bis 6,62% und für die EM-Allelrekonstruktion im Intervall von 3,33% bis 3,86%. Die empirischen Signifikanzniveaus des TDT für alle Familien und für vollständige Familien sowie der 1-TDT halten das nominelle Niveau von 5% in einem Intervall von 4,54% bis 5,56%.

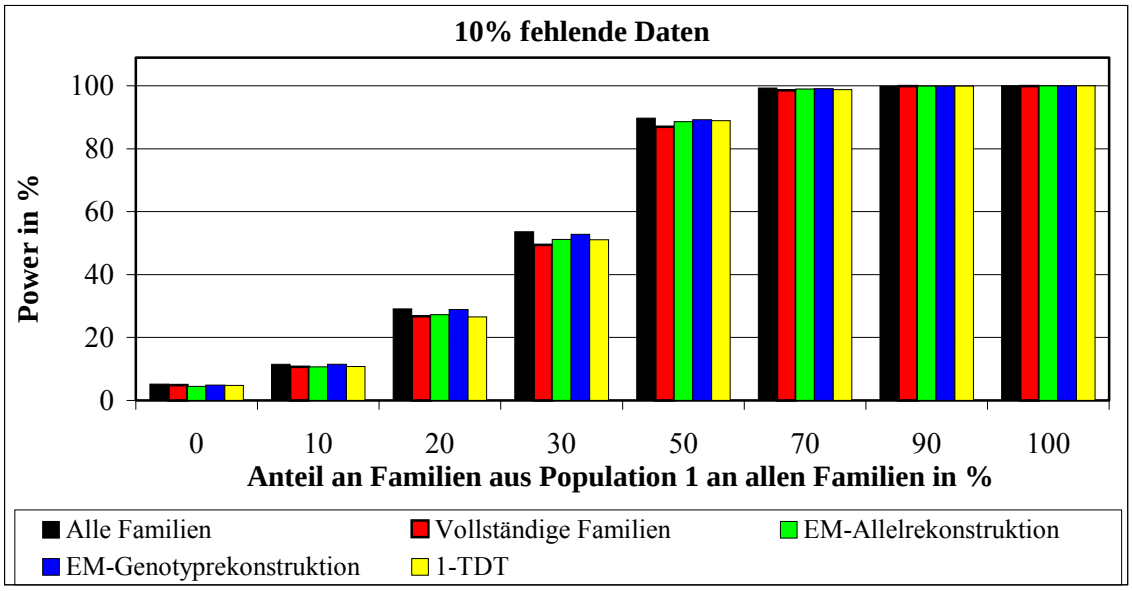
8.3 Ergebnisse unter der Alternativhypothese

MC-Simulationen wurden unter der Alternativhypothese bei Assoziation und Kopplung in Population 1 in der Variante I und II für $\theta=0$; $\delta_1=\delta_{max}$; $\delta_2=0$ sowie $\theta=0{,}01$; $\delta_1=\delta_{max}$; $\delta_2=0$ durchgeführt.

8.3.1 Rezessives Vererbungsmodell

Für das rezessive Vererbungsmodell wurden die Penetranzen $f_0=0$; $f_1=0$; $f_2=1$ festgelegt. In Abbildung 8-10 ist die Power unter der Alternativhypothese für alle Verfahren im Vergleich zu der statistischen Power bei kompletten Familien dargestellt. Bei niedrigen Allelfrequenzen für p und m besitzt der TDT schon bei einem niedrigen Anteil der Population 1 von 20% eine Power von 100%. Auch nach Löschen einiger Genotypen des Vaters sind die Powerverluste für die EM-Rekonstruktionsverfahren und den 1-TDT nicht groß. Daher wird der Powervergleich der verschiedenen Methoden bei stark unterschiedlichen Allelfrequenzen p und m gezeigt. Erst dann treten die Unterscheide untereinander deutlich hervor. In Abbildung 8-10 ist die Poweranalyse für die Parameter $p=0{,}3$; $m=0{,}1$; $\delta_1=0{,}07$; $\delta_2=0$ und $\theta=0{,}5$ dargestellt. Der Anteil fehlender Genotypen beträgt 10%.

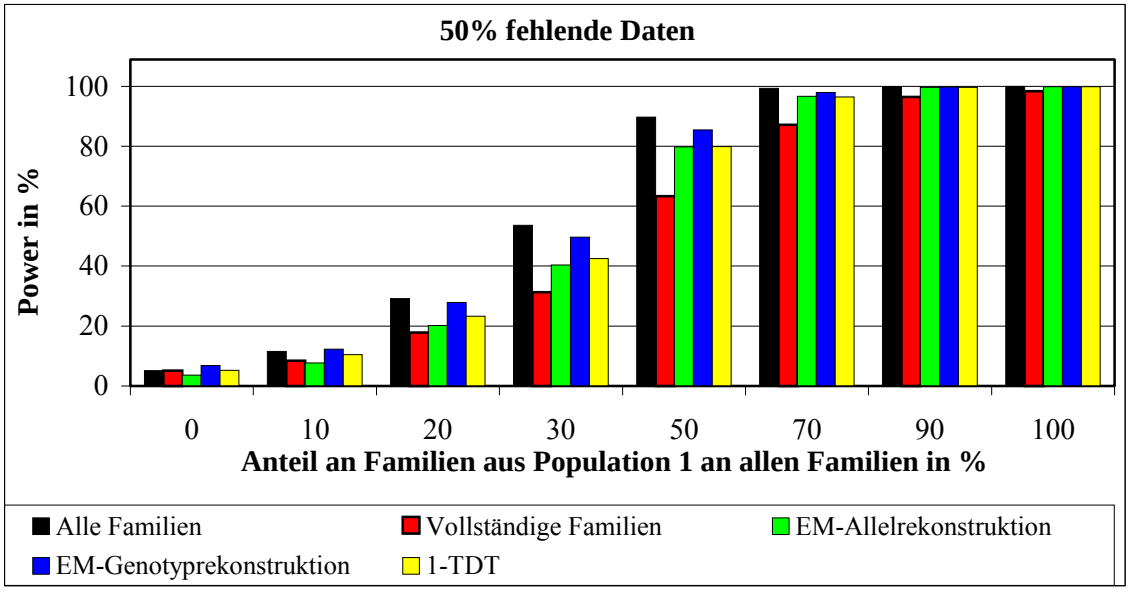
Abbildung 8-10 **Power (in Prozent)**
Transmission-Disequilibrium Test mit allen Familien, vollständigen Familien, EM-Allelrekonstruktion, EM- Genotyprekonstruktion und 1-TDT bei einem nominellen Signifikanzniveau von 5%
Globale Modellparameter: $p=0{,}3$ $m=0{,}1$ $\delta_1=0{,}07$ $\delta_2=0$ $\theta=0{,}5$
Rezessives Vererbungsmodell: $f_0=0$; $f_1=0$; $f_2=1$



Die Poweranalyse zeigt, daß bei Vorliegen der vollen Information in kompletten Familien die Wahrscheinlichkeit, eine bestehende Kopplung zu erkennen, am größten ist. Bei einem geringen Anteil von 10% der Population 1 mit vorhandener Assoziation und Kopplung ist die Wahrscheinlichkeit, ein positives Ergebnis zu erhalten, jedoch nur gering. Die Wahrscheinlichkeit, eine positive Assoziation in Population 1 zu finden, steigt in dem Maße, wie der Anteil der Population 1 zunimmt. Ab einem Anteil der Population 1 von 70% wird eine Power von 100% erreicht. Beim TDT für die vollständigen Familien, bei den beiden EM-Rekonstruktionsmethoden und dem 1-TDT sind kleine Powerverluste gegenüber dem TDT für alle Familien zu erkennen. Die Powerverluste sind jedoch in keinem Fall größer als 2,52 Prozentpunkte.

Für die gleichen Populationsparameter wird in Abbildung 8-11 die Situation für einen Anteil von 50% fehlender Genotypen illustriert.

Abbildung 8-11 **Power (in Prozent)**
Transmission-Disequilibrium Test mit allen Familien, vollständigen Familien, EM-Allelrekonstruktion, EM- Genotyprekonstruktion und 1-TDT bei einem nominellen Signifikanzniveau von 5%
Globale Modellparameter: $p=0,3$ $m=0,1$ $\delta_1=0,07$ $\delta_2=0$ $\theta=0,5$
Rezessives Vererbungsmodell: $f_0=0$; $f_1=0$; $f_2=1$



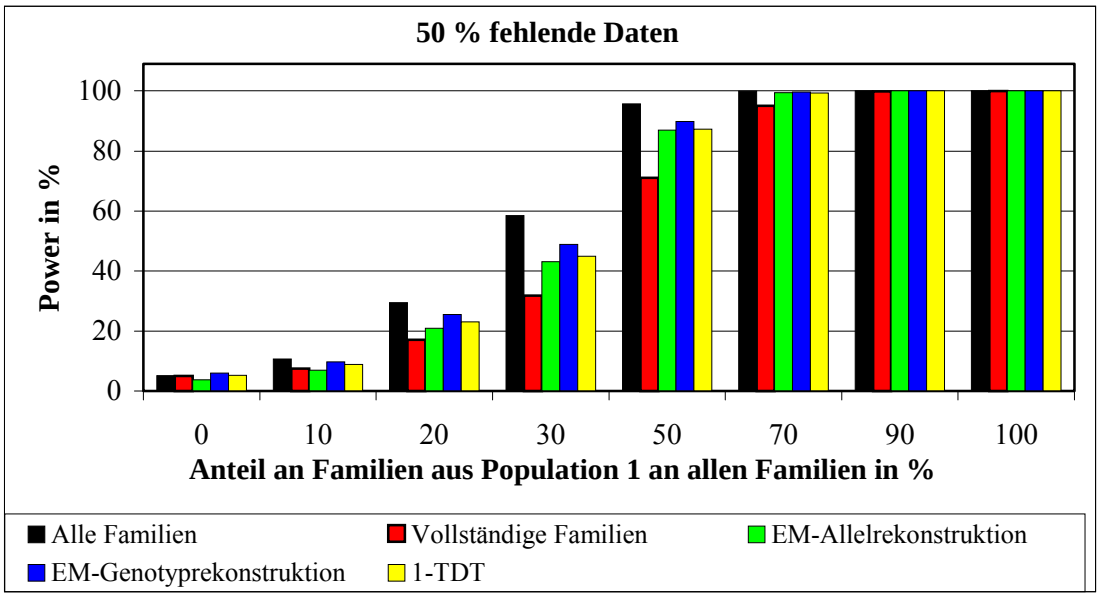
Wenn für die Berechnung der TDT-Statistik nur die vollständigen Familien verwendet werden, ist in Abbildung 8-11 ein deutlicher Verlust an Power gegenüber dem TDT für die kompletten Familien bei 50% fehlenden Daten erkennbar. Der maximale Powerverlust beträgt dabei 26,46 Prozentpunkte. Die Methode der EM-Rekonstruktion und der 1-TDT führen zu einem erkennbaren Gewinn an Power im Vergleich zur

ausschließlichen Verwendung der vollständigen Familien. Die EM-Genotyprekonstruktion gewinnt die meiste Power zurück. Der maximale Powerverlust beträgt nach der EM-Genotyprekonstruktion 4,29 Prozentpunkte. Die EM-Allelrekonstruktion und der 1-TDT führen auch zu einem Powergewinn, aber erkennbar weniger im Vergleich zur EM-Genotyprekonstruktion. Der Powerverlust beträgt nach der EM-Allelrekonstruktion maximal 13,24 Prozentpunkte und beim 1-TDT 11,16 Prozentpunkte.

8.3.2 Dominantes Vererbungsmodell

In Abbildung 8-12 ist die Power für alle Auswertungen unter einem dominanten Vererbungsmodell mit den Penetranzen f_0=0, f_1=1 und f_2=1 dargestellt. Der Poweranalyse liegen die Parameter p=0,05; m=0,3; δ_1=0,035; δ_2=0 und θ=0,5 zugrunde. Der Anteil fehlender Daten beträgt 50%.

Abbildung 8-12 **Power (in Prozent)**
Transmission-Disequilibrium Test mit allen Familien, vollständigen Familien, EM-Allelrekonstruktion, EM-Genotyprekonstruktion und 1-TDT bei einem nominellen Signifikanzniveau von 5%
Globale Modellparameter: p=0,05 m=0,3 δ_1=0,035 δ_2=0 θ=0,5
Dominantes Vererbungsmodell: f_0=0; f_1=1; f_2=1

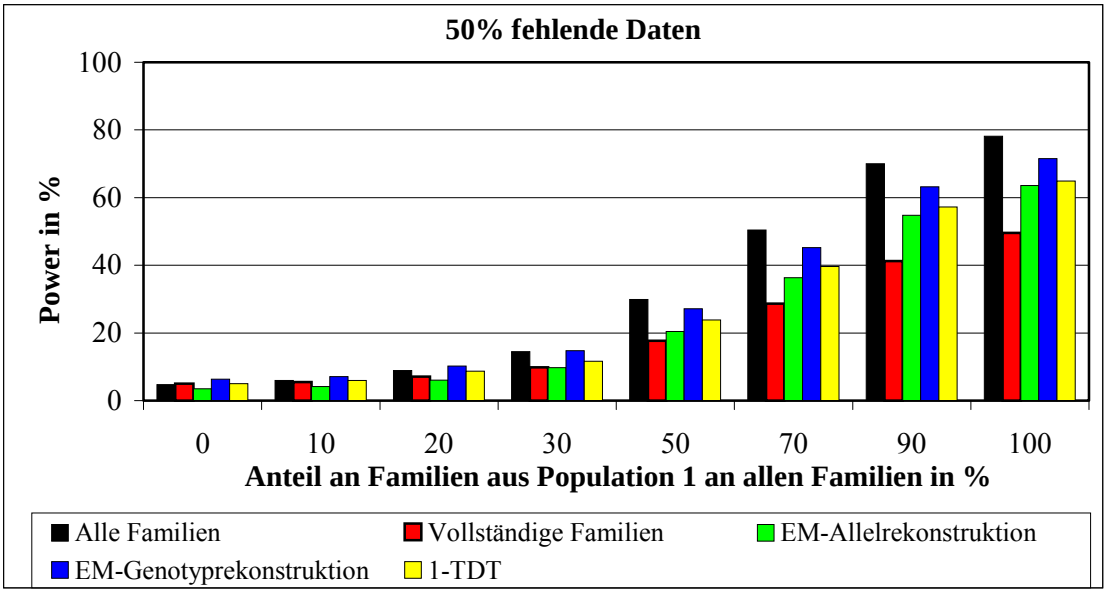


Der TDT für alle Familien erreicht ab einem Anteil der Population 1 von 70% eine Power von 100%. Vergleichbar zu den Ergebnissen bei der rezessiven Vererbung wird der größte Powerverlust bei dem TDT für die vollständigen Familien beobachtet. Die Methoden der EM-Rekonstruktion und der 1-TDT können bei einem Anteil der Population 1 von 50% mehr als die Hälfte des Powerverlustes des TDT für vollständige

Familien zurückgewinnen. Die höchsten Powerwerte erreicht auch bei diesem Modell die EM-Genotyprekonstruktion. Der maximale Powerverlust beträgt nur 4,29 Prozentpunkte. Die maximalen Powerverluste sind bei der EM-Allelrekonstruktion 13,24 Prozentpunkte und beim 1-TDT 11,16 Prozentpunkte.

Eine zweite Poweranalyse wird in der folgenden Abbildung gezeigt für die Parameter p=0,3; m=0,1; δ_1=0,07; δ_2=0 und θ=0,5 sowie 50% fehlender Genotypen.

Abbildung 8-13 **Power (in Prozent)**
Transmission-Disequilibrium Test mit allen Familien, vollständigen Familien, EM-Allelrekonstruktion, EM-Genotyprekonstruktion und 1-TDT bei einem nominellen Signifikanzniveau von 5%
Globale Modellparameter: p=0,3 m=0,1 δ_1=0,07 δ_2=0 θ=0,5
Dominantes Vererbungsmodell: f_0=0; f_1=1; f_2=1

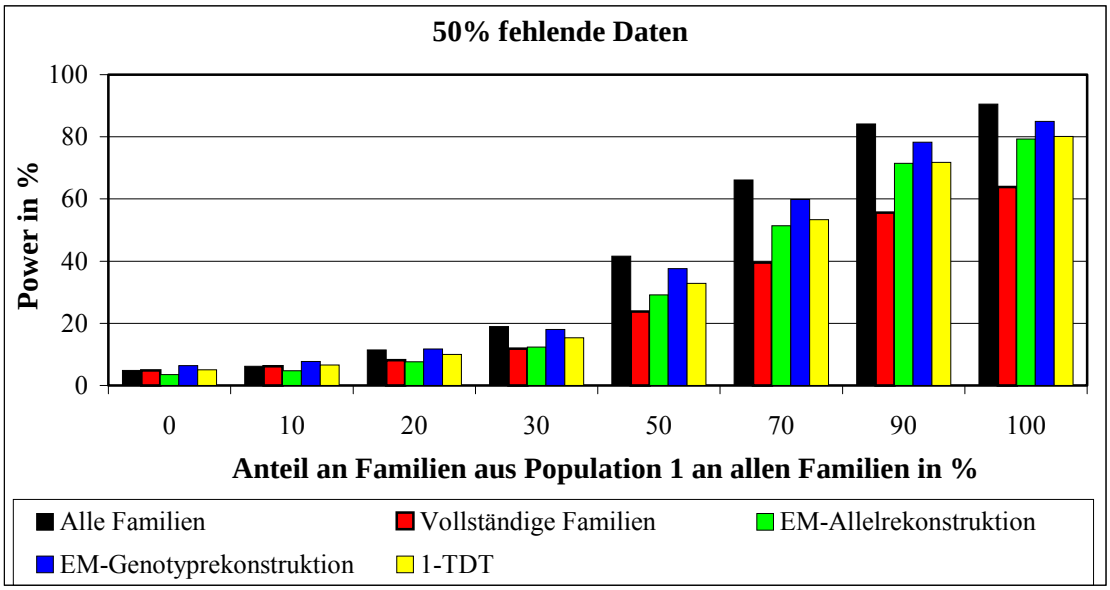


Im Vergleich mit Abbildung 8-11, in der die gleichen Parameter für ein rezessives Modell gewählt wurden, erreicht der TDT für die kompletten Familien auch bei einem Anteil der Population 1 von 100% nicht die Power von 100%, sondern nur 78,15%. Der Powerverlust für die vollständigen Familien ist in diesem Fall dagegen mit bis zu 28,77 Prozentpunkten sehr groß. Die anderen TDT-Teststatistiken können auch hier erheblich an Power zurückgewinnen, wobei der Gewinn bei der EM-Genotyprekonstruktion erneut am deutlichsten ist.

8.3.3 Additives Vererbungsmodell

In Abbildung 8-14 sind die Powerwerte unter einem additiven Vererbungsmodell mit den Penetranzen f_0=0, f_1=0,5 und f_2=1 und für die Parameter p=0,3; m=0,1; δ_1=0,07; δ_2=0; θ=0,5 bei 50% fehlenden Daten dargestellt.

Abbildung 8-14 **Power (in Prozent)**
Transmission-Disequilibrium Test mit allen Familien, vollständigen Familien, EM-Allelrekonstruktion, EM- Genotyprekonstruktion und 1-TDT bei einem nominellen Signifikanzniveau von 5%
Globale Modellparameter: p=0,3 m=0,1 δ_1=0,07 δ_2=0 θ=0,5
Additives Vererbungsmodell: f_0=0; f_1=0,5; f_2=1

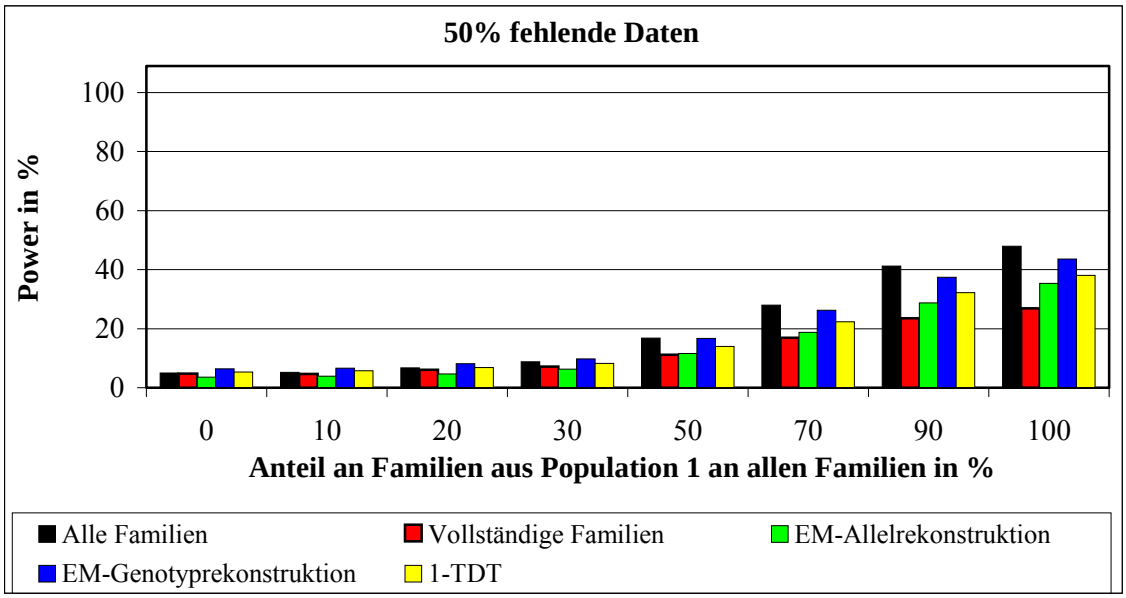


Wie schon bei den Powerwerten unter dem dominanten Modell bei gleichen Parametern überbieten bei allen Graden von Populationsstratifikation weniger als 100% der Testwerte einen Wert von 3,8414588. Die maximale Power ist mit 90,49% unter diesem Vererbungsmodell jedoch höher als unter dem dominanten Modell mit 78,15%. Der maximale Powerverlust beim TDT für vollständige Familien beträgt 28,60 Prozentpunkte, bei der EM-Allelrekonstruktion 14,78 Prozentpunkte, bei der EM-Genotyprekonstruktion 6,32 Prozentpunkte und beim 1-TDT 12,75 Prozentpunkte.

8.3.4 Additives Vererbungsmodell mit Phänokopie

Abbildung 8-9 zeigt die Power der verschiedenen Teststatistiken unter einem additiven Vererbungsmodell mit Phänokopie mit den Penetranzen f_0=0,1; f_1=0,5; f_2=0,5 und den Parametern p=0,3; m=0,1; δ_1=0,07; δ_2=0; θ=0,5 bei 50% fehlenden Daten.

Abbildung 8-15 **Power (in Prozent)**
Transmission-Disequilibrium Test mit allen Familien, vollständigen Familien, EM-Allelrekonstruktion, EM- Genotyprekonstruktion und 1-TDT bei einem nominellen Signifikanzniveau von 5%
Globale Modellparameter: p=0,3 m=0,1 δ_1=0,07 δ_2=0 θ=0,5
Additives Vererbungsmodell mit Phänokopie: f_0=0,1; f_1=0,5; f_2=0,5



Verglichen mit den Poweranalysen unter den zuvor dargestellten Vererbungsmodellen werden in dem Modell aus Abbildung 8-15 auch für den TDT bei kompletten Familien nur geringe Powerwerte erreicht. Bei einem Anteil der Population 1 von 100% wird in 47,96% der Fälle eine bestehende Assoziation und Kopplung erkannt. Bei der Anwendung der TDT-Statistik nur auf die vollständigen Familien bei einem Anteil der Population 1 von 100% liegt die maximale Power bei 26,85%. Werden die Informationen der Paare berücksichtigt, kann die Power bei der EM-Allelrekonstruktion auf 35,34% gesteigert werden. Bei der EM-Genotyprekonstruktion wird der größte Powergewinn mit einem Anstieg auf 43,56% beobachtet. Die Powergewinn des 1-TDT liegt mit einem erreichten Power von 38,07% ein wenig höher als bei der EM-Allelrekonstruktion. Dieses Verhältnis zueinander besteht auch bei einer Populationsstratifikation von 70% und 90%.

8.4 Anwendung auf Realdaten

Zur Demonstration der Anwendung der EM-Rekonstruktionsmethoden und des 1-TDT wurde ein Reanalyse von bereits partiell veröffentlichten Daten mit Kindern und Erwachsenen mit Anorexia nervosa und Adipositas per magna durchgeführt. Die

Genotypen der Testpersonen und soweit vorhanden auch die Genotypen der Eltern wurden an zwei verschiedenen Markerloci typisiert.

Der erste Markerlocus kodiert für den β-adrenergen Rezeptor mit den Allelen β_1, β_2 und β_3. Durch Bindung von Katecholaminen insbesondere an den β_3-adrenergen Rezeptor wird die hormonsensitive Lipase aktiviert, das Schlüsselenzym der Lipolyse. Eine Punktmutation von Tryptophan zu Arginin (Trp64Arg) wurde zuvor als Risikofaktor für Übergewicht beschrieben (HINNEY *et al*., 1997b). Es konnten 99 untergewichtigen, 80 normalgewichtigen und 238 übergewichtigen Kindern und Erwachsenen, sowie 84 Patienten mit Anorexia nervosa am β-adrenergen Rezeptor typisiert werden. Bei 80 Familien mit übergewichtigen Patienten und 52 Familien mit Patienten mit Anorexia nervosa wurde der TDT angewandt. Dem Trp64Arg Polymorphismus des β_3-adrenergen Rezeptor konnte keine wesentliche Rolle bei der Regulation des Körpergewichtes zugeordnet werden.

Das zweite Markerallel liegt in der Genregion des Serotonin-Transporter Proteins (5-HTT). Serotonin ist ein Neurotransmitter mit vielfältigen Funktionen insbesondere im Energie- und Hormonstoffwechsel. 5-HTT reguliert die Wiederaufnahme von extrazellulärem Serotonin ins ZNS. Einem häufigen Polymorphismus von repetitiven Einheiten in der Serotonin-Transporter-Region (5-HTTLRP) eine Rolle in der Entstehung von Eßstörungen zugeschrieben (HINNEY *et al.*, 1997a). Alle Studienkollektive wurden mit der Methode der Polymerase-Ketten-Reaktion (PCR) typisiert. Es konnten 385 übergewichtige Kinder, Jugendliche und Erwachsene, 112 untergewichtigen Patienten und 55 Patienten mit Anorexia nervosa man Serotonin-Transporter-Region typisiert werden. Ein Vergleich der Allelfrequenzen zwischen übergewichtigen und untergewichtigen Patienten erbrachte keinen Hinweis auf eine Rolle des 5-HTTLRP Polymorphismus in der Serotonin-Transporter-Region bei der Regulation des Körpergewichtes.

Die Datensätze für die zwei Markerloci wurden in Gruppen mit vollständigen Familientrios und Paaren aus Kind und einem Elternteil aufgeteilt. Die Rekonstruktion der fehlenden Daten erfolgt durch die Prozeduren `ESTIMm2`, `ESTIMg2`, `REKONSTR2` und `REKONSTR4`. Die Berechnung der TDT-Statistik durch die Prozedur `TDTSTAT` wurde jeweils für die vollständigen Familien und die vollständigen Familien mit den rekonstruierten Trios aus den Paaren sowohl nach der

EM-Allelrekonstruktion als auch nach der EM-Genotyprekonstruktion durchgeführt. Außerdem wurde für die gleichen Datender 1-TDT mit der Prozedur `REKONSTR7` berechnet.

8.4.1 Anorexia nervosa und β3-adrenerge Rezeptor-Polymorphismus

In eine Studie an der Universität Marburg wurden 99 Studenten mit einem BMI ≤ 15. Altersperzentile, Fehlen einer somatischen Erkrankung, keine Erkrankung an Anorexia nervosa in der Vorgeschichte und einem Zigarettenkonsum ≤ 10/Tag aufgenommen und wurden am β3-adrenerge Rezeptor-Polymorphismus typisiert. Außerdem wurden 84 Patienten mit Anorexia nervosa, die die DSM-IV Kriterien erfüllen, aus der Kinder- und Jugendpsychiatrien der Universitäten Köln, Frankfurt, Mannheim und Marburg sowie aus der Inneren Medizin der Universität Heidelberg typisiert. In 52 Fällen konnten zusätzlich beide Eltern typisiert werden und in 18 Fälle war nur ein Elternteil verfügbar.

Die Information der vollständigen Trios ist in einer T/NT-Tabelle dargestellt. Aus dieser Tabelle wurde der p-Wert der TDT-Statistik für die kompletten Familien berechnet.

Tabelle 8-6 Vollständig typisierte Familientrios am β3-adrenergen Rezeptor-Polymorphismus bei Anorexia nervosa

Transmittiert	**Nicht transmittiert**	
	1	2
1	0	4
2	8	92
TDT p=0,248		

Der p-Wert überschreitet das Signifikanzniveau von 5%, so daß in diesem Fällen die Nullhypothese beibehalten wird. Es wird daher davon ausgegangen, daß die Trp64Arg-Variante des β3-adrenerge Rezeptors mit Anorexia nervosa nicht assoziiert oder gekoppelt ist.

Die Informationen der Paare wurden in die folgende Tabelle eingetragen. Mit den Informationen aus Tabelle 8-6 und Tabelle 8-7 kann der 1-TDT angewandt werden.

Tabelle 8-7 Genotypinformationen der Paare aus einem erkranktem Kind und einem Elternteil am β3-adrenergen Rezeptor-Polymorphismus bei Anorexia nervosa

Kind	**Elternteil**		
	11	12	22
11	0	0	0
12	0	0	3
22	0	1	14
1-TDT		p=0,617	

Auch bei der Berechnung der Teststatistik des 1-TDT für die 18 Paare wurde das nominelle Niveau von 5% nicht unterboten.

Auf die Informationen der Paare wurde weiterhin die Methode der EM-Allel- und EM-Genotyprekonstruktion angewendet. Nach erfolgreicher Rekonstruktion wurden bei dieser Methode relativen Einträge in die T/NT-Tabelle vorgenommen. Das Ergebnis ist in der folgenden Tabelle dargestellt.

Tabelle 8-8 Rekonstruierte Trios aus Paaren am β3-adrenergen Rezeptor-Polymorphismus bei Anorexia nervosa

EM-Allelrekonstruktion			**EM-Genotyprekonstruktion**		
Transmittiert	Nicht transmittiert		Transmittiert	Nicht transmittiert	
	1	2		1	2
1	0,13	2,87	1	0	4,00
2	1,63	31,37	2	1,22	30,78

Die Einträge aus Tabelle 8-6 und Tabelle 8-8 wurden nun kombiniert und in einer Gesamttabelle addiert. Aus dieser Tabelle wurde dann der p-Wert der TDT-Statistik für die kombinierten Daten berechnet.

Tabelle 8-9 Vollständige Familien und rekonstruierte Paare am β3-adrenerge Rezeptor-Polymorphismus bei Anorexia nervosa

EM-Allelrekonstruktion			**EM-Genotyprekonstruktion**		
Transmittiert	Nicht transmittiert		Transmittiert	Nicht transmittiert	
	1	2		1	2
1	0,13	6,87	1	0	8
2	9,63	123,37	2	9,22	122,78
TDT p=0,497			TDT p=0,789		

In keinem der Fälle wird das nominelle Niveau von 5% überschritten. Es wird also auch nach der Rekonstruktionen keine Assoziation und Kopplung entdeckt.

8.4.2 Anorexia nervosa und Serotonin-Transporter-Polymorphismus

Die Testpersonen mit Anorexia nervosa stammen aus einer Gruppe von 146 untergewichtigen Studenten der Universität Marburg mit einem BMI ≤ 15. Altersperzentile, Fehlen einer somatischen Erkrankung, keine Erkrankung an Anorexia nervosa in der Vorgeschichte und einem Zigarettenkonsum ≤ 10/Tag sowie aus einer Gruppe von 96 Patienten der Kinder- und Jugendpsychiatrien Köln, Frankfurt, Mannheim und Marburg und der Inneren Medizin der Universität Heidelberg, die die DSM-IV Kriterien der Anorexia nervosa erfüllen. Unter diesen Testpersonen konnten 54 vollständige Familientrios und 18 Paare am Serotonin-Transporter-Polymorphismus typisiert werden. Bei den Trios und die Paare wurde die Methode der EM-Rekonstruktion und der 1-TDT angewandt. Die vollständigen Trios wurden in die T/NT-Tabelle eingetragen (Tabelle 8-10).

Tabelle 8-10 Vollständige Familientrios am Serotonin-Transporter-Polymorphismus bei Anorexia nervosa

Transmittiert	**Nicht transmittiert**	
	1	2
1	33	27
2	26	22
TDT p=0,891		

Die Paare wurden in Tabelle 8-11 eingetragen und der 1-TDT berechnet.

Tabelle 8-11 Genotypinformationen der Paare aus einem erkranktem Kind und einem Elternteil am Serotonin-Transporter-Polymorphismus bei Anorexia nervosa

Kind	**Elternteil**		
	11	12	22
11	1	3	0
12	0	3	5
22	0	4	1
1-TDT p=0,317			

Nach Rekonstruktion der EM-Allelrekonstruktion bei den Paare und Kombination der Ergebnisse ergab sich ein p-Wert von p=0,598. Bei der EM-Genotyprekonstruktion ergab sich nach der Rekonstruktion bei Paare p=0,224. Alle p-Werte bleiben oberhalb des nominellen Signifikanzniveaus von 5%. Weder die EM-Rekonstruktionsmethoden

noch der 1-TDT geben einen Hinweis auf das Vorliegen von Assoziation und Kopplung.

8.4.3 Adipositas per magna und β3-adrenergen Rezeptor-Polymorphismus

Am β3-adrenergen Rezeptor-Polymorphismus wurden 166 extrem übergewichtige Kindern und Erwachsene, deren BMI in den meisten Fällen die 99. Perzentile übertreffen, aus dem Kinderkrankenhaus Murnau typisiert. Eine zweite Gruppe mit 72 weniger extrem übergewichtigen Kindern wurde durch Schulärzte, Kinderärzte und Zeitungsanzeigen rekrutiert. Insgesamt konnten zusätzlich 65 Eltern und 32 Paare am β3-adrenergen Rezeptor-Polymorphismus typisiert. Die EM-Rekonstruktionen bei den Paaren erfolgt nach den gleichen Prozeduren wie bei den Daten zu Anorexia nervosa. Die TDT-Statistik wird ebenfalls für die kompletten Familien und für komplette Trios nach erfolgreichen Rekonstruktionen berechnet. Außerdem wurde die Teststatistik des 1-TDT berechnet. Der p-Wert der TDT-Statistik bei den vollständigen Trios beträgt p=0,796.

Tabelle 8-12 Berechnete p-Werte des TDT und des kombinierten 1-TDT am β3-adrenergen Rezeptor-Polymorphismus bei Adipositas per magna

Test	Anzahl der Familien	p-Wert
TDT	65	0,796
1-TDT	97	0,108
EM-TDT Allelrekonstruktion	97	0,789
EM-TDT Genotyprekonstruktion	97	0,622

Die p-Werte bleiben in allen Fällen über dem nominellen Signifikanzniveau von 5%.

8.4.4 Adipositas per magna und Serotonin-Transporter-Polymorphismus

Die Testpersonen mit Adipositas per magna, die am Serotonin-Transporter-Polymorphismus typisiert wurden, stammen aus drei Studien. Die erste Studie umfaßt 229 Kindern und Erwachsener aus einer deutschen Population, die in den meisten Fällen die 100. BMI-Perzentile erreichen. Die Testpersonen der zweiten Studie stammen aus

einer Gruppe von 48 übergewichtigen Studenten der Universität Marburg mit einem BMI ≥ 90. Perzentile, Fehlen einer somatischen Erkrankung, keine Erkrankung an Anorexia nervosa in der Vorgeschichte und einem Zigarettenkonsum ≤ 10/Tag. Die dritte Studie enthält 88 Personen mit einem BMI ≥ 35 kg/m^2 aus Klinik für Innere Medizin des Universitätskrankenhaus Eppendorf, Hamburg. Insgesamt konnten 94 vollständige Familientrios und 38 Paar typisieren werden. Der p-Wert der TDT-Statistik bei den vollständigen Familien beträgt p=0,345.

Tabelle 8-13 Berechnete p-Werte des TDT und des 1-TDT am Serotonin-Transporter-Polymorphismus bei Adipositas per magna

Test	Anzahl der Familien	p-Wert
TDT	94	0,345
1-TDT	132	0,499
EM-TDT Allelrekonstruktion	132	0,373
EM-TDT Genotyprekonstruktion	132	0,987

Das nominelle Signifikanzniveau von 5% wird auch in diesem Fall immer unterboten.

9 Diskussion

Aufgabe der genetischen Epidemiologie ist die Untersuchung von genetischen Faktoren und Umweltfaktoren, die auf die Entstehung und den Verlauf einer Krankheit einen Einfluß haben. Mit dem Wissen über ein Krankheitsgen verspricht man sich ein besseres Verständnis der Pathogenese einer Krankheit. Daraus kann möglicherweise eine bessere Diagnose gestellt und später eine spezifische und kausale Therapie entwickelt werden. Erfolgreiche Beispiele hierfür sind eine atypisch verlaufende Mukoviszidose (OMIM, 219700) sowie die Hämochromatose (OMIM, 235200). Von Vorteil bei der Erforschung dieser Krankheiten sind Vererbungsmodelle, die streng den Mendelschen Regeln folgen. Bei diesen Krankheiten ist es vergleichsweise einfach, den bzw. die unbekannten Genloci zu finden. So konnten schon mehr als 500 Krankheitsgenen bis Ende 1995 eine chromosomale Region zugeordnet werden und bei 60 Krankheitsgenen ist exakte Position angegeben werden (LANDER & KRUGLYAK, 1995). Bedeutend schwieriger ist die Situation bei komplexen Krankheiten wie z.B. Diabetes mellitus Typ II. Bei Diabetes mellitus Typ II gibt es viele verdächtige Genloci, sogenannte Kandidatengene, denen ein erhöhte Suszeptibilität zugeschrieben wird (OMIM, 125853). Weiterhin gibt es eine Reihe von Umweltfaktoren wie z.B. Virusinfektionen, die teilweise sogar eine größere Bedeutung als genetische Faktoren haben. Zur Analyse von Kandidatengenen und zur Feinkartierung chromosomaler Regionen wurden in den letzten Jahren eine Reihe von Studiendesigns entwickelt.

In der genetischen Epidemiologie werden bei der Suche nach funktionsrelevanten DNA-Varianten zwei verschiedene Ansätze verwendet, das sind Kopplung und Assoziation. Bei messen auf unterschiedliche Weise eine Abweichung von den Mendelschen Regeln bei unabhängiger Vererbung. Bei Kopplung wird innerhalb von Familien ein Phänotyp und ein genetischer Marker überzufällig häufig gemeinsam weitervererbt. Dabei wird angenommen, daß die DNA-Sequenz, die dem Phänotypen zugrunde liegt, so nah neben dem genetischen Marker liegt, daß beide Genloci nur selten durch Rekombination voneinander getrennt werden. Dagegen spricht man von Assoziation, wenn innerhalb einer Population ein spezifisches Allel eines genetischen Markers überzufällig bei erkrankten Personen vorkommt. Kopplung und Assoziation können auch gemeinsam vorkommen. In diesem Fall werden ein spezifisches Allel und eine Krankheit in vielen Familien zusammen weitervererbt.

Da mit steigendem Abstand zwischen Krankheitslocus und Markerlocus die Wahrscheinlichkeit steigt, daß beide Loci in irgendeiner Generation durch Rekombination getrennt werden und so die Krankheit mit verschiedenen Allelen einhergeht, sollte für Assoziationsstudien der Abstand zwischen Krankheitslocus und Markerlocus nicht mehr als 1 cM betragen. Bei komplexen Krankheiten wird in Kopplungsstudien ein sogenannter Genomscan mit üblicherweise 300-500 Markerloci, in größeren Abständen über das gesamte Genom verteilt sind, durchgeführt (OTT, 1996). Ist durch Kopplungsstudien dann die chromosomale Region des Krankheitslocus gefunden, werden zur Untersuchung von Kandidatengene oder zur Feinkartierung der physikalische chromosomale Region Assoziationsstudie eingesetzt (STRACHAN & READ, 1999). Assoziationsstudien besitzen bei der Identifikation von Allelen, die einen geringen Einfluß auf eine Krankheit haben, eine höhere Power als Kopplungsstudien (RISCH & MERIKANGAS, 1996). Erstmalig wurde dieses Vorgehen von HORIKAWA *et al.* (2000) erfolgreich bei CAPN10 und Diabetes mellitus Typ II angewendet. Nachdem HANIS *et al.* (1996) durch Kopplungsanalysen unter mexikanischstämmigen US-Amerikanern einen verdächtigen Genlocus, bezeichnete als NIDDM1, im Chromosomenband 2q37.3 gefunden hatten, konnten HORIKAWA *et al.* (2000) durch Feinkartierung den verdächtigen Genabschnitt von 1,7 Mb auf 240 kb eingrenzen. Nach Positionsklonierung wurde das Gen CAPN10 gefunden, das eine Assoziation bei mexikanischstämmigen US-Amerikanern, Nordeuropäern in Finnland mit Diabetes mellitus Typ II zeigte.

Die einfachste Variante auf Assoziation zu testen ist die klassische Fall-Kontroll Studie, bei der verglichen wird, ob ein spezifisches Allel überzufällig häufig in einer Gruppe von kranken Personen vorkommt im Vergleich zu einer Gruppe mit gesunden Personen. Dabei entsteht Assoziation, wenn ein genetischer Marker in enger Nachbarschaft bzw. innerhalb des Krankheitsgenlocus liegt oder wenn der Markerlocus identisch mit dem Krankheitslocus ist. Eine positive Assoziation jedoch stets äußerst kritisch interpretiert werden. Denn es ist auch möglich, daß die positive Assoziation auf Scheinassoziation beruht. Scheinassoziation, die entsteht, wenn eine Population aus unterschiedliche Subpopulationen mit verschiedenen Frequenzen für das Krankheitsgen und die Markerallele besteht. Mit Scheinassoziation muß daher gerechnet werden, wenn Fälle und Kontrolle nicht sorgfältig genug ausgesucht werden (MORTON & COLLINS, 1998).

Zur Umgehung des Problems von Scheinassoziation wurden familienbasierte Assoziationsstudien entwickelt. Das sogenannte Haplotype-Relative Risk (FALK & RUBINSTEIN, 1987) bildet aus nicht-transmittierten elterlichen Allelen eine fiktive Kontrollgruppe und verzichtet damit auf eine reale Kontrollgruppe. Vorteil diese Vorgehens ist, daß die möglichen Probleme durch eine inadäquate Auswahl der Kontrollpersonen vermieden werden. Denn die fiktive Kontrollgruppe wird exakt aus der selben Gruppe entnommen, wie die Fallgruppe. Dadurch wird Confounding bei ethnischer Heterogenität vermieden. Außerdem erfolgt die Typisierung immer im selben Labor und Fehltypisierungen können durch Vergleich der Segregationsmuster bei Mendelscher Vererbung erkannt werden. Dennoch ist das HRR nur ein Test auf Assoziation und kann nicht auf Kopplung testen, da das HRR zusätzlich zur Assoziationsinformation des nicht transmittierten elterlichen Allels nicht die Kopplungsinformation nutzt, die in der Segregation der elterlichen Allele auf das erkrankte Kind enthalten sind.

Eine Weiterentwicklung ist der Transmission-Disequilibrium Test (TDT) von SPIELMAN *et al.* (1993). Der TDT testet, ob heterozygote Eltern an einem Markerlocus ein spezifisches Allel bevorzugt an ihr erkranktes Kind weitervererben. Dabei ist es irrelevant, ob die Eltern auch erkrankt sind oder nicht. Dabei sei a die Anzahl der Fälle bei denen ein heterozygotes Elternteil das Allel 1 vererbt und b die Anzahl der Fälle bei denen das Allel 2 vererbt wird. Die Teststatistik des TDT ist $(a-b)/(a+b)^2$ und hat eine χ^2-Verteilung mit einem Freiheitsgrad. Die große Bedeutung des TDT beruht auf der Eigenschaft, nicht nur die Assoziationsinformation, sondern auch die Kopplungsinformation zu nutzen, die in der Vererbung der elterlichen Allel an das erkrankte Kind enthalten sind. Damit kann der TDT Kopplung entdecken, wenn gleichzeitig Assoziation besteht. Umgekehrt kann er Assoziation nur dann entdecken, wenn gleichzeitig Kopplung vorhanden ist (EWENS & SPIELMAN, 1995). Daher ist der TDT auch ein kombinierter Kopplungs- und Assoziationstest. Aufgrund der einfachen Handhabung, Robustheit und seiner relativ großen Power ist der TDT in der Praxis häufig angewandt und erweitert worden.

Eine Grundbedingung für den TDT ist, daß beide Eltern am Markerlocus typisiert sein müssen. Daraus ergibt sich ein Problem bei Erkrankungen mit einem späten Manifestationsalter, denn dann ist es häufig schwierig bis unmöglich, beide Elternteile gleichzeitig zu genotypisieren. Formal kann zwar in bestimmten Fällen die

Transmissionsinformation von Elter-Kind Paare für die TDT-Statistik genutzt werden, aber daraus resultiert ein Bias. Die Transmission eines bestimmtes Allel wird bei einem heterozygoten Elterteil nicht nur bei Assoziation und Kopplung, sondern dann häufiger beobachtet, wenn das Allel häufiger bei den Eltern vorkommt (CURTIS & SHAM, 1995). Zur Umgehung des Problems wurden der Sib-TDT (SPIELMAN & EWENS, 1998) und der SDT (HORVATH & LAIRD, 1998) vorgeschlagen. So benutzt man beim Sib-TDT Geschwister als Kontrollen und untersucht, ähnlich einer gematchten Fall-Kontrollstudie, ob ein erkranktes Geschwisterkind überzufällig häufig ein spezifisches Markerallel besitzt als sein nicht erkranktes Geschwister. Die Frequenz des Markerallels bei erkrankten Kindern wird mit der Frequenz des Markerallels bei gesunden Kindern verglichen. Der SDT hat einen ähnlichen Ansatz, vergleicht jedoch die Frequenzen des Markerallels in jeder Familie einzeln und ist auch auf multiallelische Marker anwendbar. Dieser Ansatz macht es möglich, den Sib-TDT und den SDT auch bei Erkrankungen mit einem hohen Manifestationsalter wie Diabetes mellitus Typ II, kardiovaskulären Erkrankungen oder Demenz vom Alzheimer Typ anzuwenden. Dennoch sollte dieses Rekrutierungsschema bei der Anwendung in der Praxis sorgfältig geprüft werden, da dieses Design bei einem rezessiven Modell nur eine geringe Power besitzt (ZIEGLER & HEBEBRAND, 1998). Zwar scheint der Ansatz bei Geschwisterpaaren mit extremer Diskordanz in Bezug auf den Phänotyp der Erkrankung besonders praktikabel, aber kulturelle und soziale Faktoren können dieses Phänomen verzerren. So fanden ZIEGLER & HEBEBRAND (1998) in einer Studie heraus, daß sich bei einer Jugendlichen mit Anorexia nervosa das extreme Untergewicht vermutlich als Reaktion auf die Adipositas der Schwester entwickelt hatte. Weiterhin kann Diskordanz entstehen, wenn bei altersabhängigen Penetranzen jüngere Geschwister als Kontrollen verwendet werden. Darüber hinaus gibt es Probleme bei Krankheiten mit reduzierten Penetranzen, wie z.B. familiärem Brustkrebs oder Retinopathia pigmentosa. Allerdings sind der Sib-TDT und der SDT nur anwendbar, wenn bei fehlenden elterlichen Daten Geschwister vorhanden sind. Sind allerdings ausschließlich Elter-Kind Paare vorhanden, sind der Sib-TDT und der SDT nicht anwendbar.

Alternativ wurde von KNAPP (1999) der sogenannte „reconstruction combined TDT“ (RC-TDT) vorgeschlagen. Die Idee das RC-TDT ist die Rekonstruktion der fehlenden Genotypen, wenn sowohl erkrankte als auch gesunde Geschwisterkinder an einem multiallelischen Marker typisiert sind. Bei einem hohen Maß an

Markerpolymorphismus und einer hohen Fallzahl hat der Ansatz des RC-TDT eine höhere Power als der Sib-TDT, da in diesem Fall die Wahrscheinlichkeit steigt, den fehlenden Genotypen aus den Genotypinformationen der Kinder zu bestimmen. Da für den RC-TDT ein multiallelischer Marker und gesunde Geschwisterkinder notwendig sind, kann er nicht auf die bisher beschriebenen Elter-Kind Paaren mit einem biallelische Marker angewandt werden.

Ein ähnliches Verfahren der Rekonstruktion der fehlenden Daten wurden von MARTIN *et al.* (1998) vorgeschlagen. Die sogenannte Parental Genotype Reconstruction (PRG) basiert auf dem Haplotype Relative Risk (HRR) und rekonstruiert die fehlenden Daten unter Verwendung der Genotypinformation der Geschwister. Mit geschätzten Allelfrequenzen werden internen Kontrollen unter der Annahme keiner Assoziation $\delta=0$ gebildet, um keine Verzerrung bei der Schätzung zu erzeugen und den Fehler 1. Art nicht zu erhöhen. In Simulationen wurde gezeigt, daß der Fehler 1. Art nur sehr gering steigt und trotz der Annahme von $\delta=0$ die Power steigt, mit der Assoziation nachzuweisen ist. Nachteil der Methode bleibt jedoch die zusätzlich notwendige Typisierung von Geschwisterkindern. Von WEINBERG (1999) wurde ein Verfahren zur Schätzung der fehlenden Genotypen auf der Basis der Maximum-Likelihood-Methode vorgestellt. Das Schätzprinzip des likelihood-ratio test (LRT) für die fehlenden Genotypen basiert auf der Maximum-Likelihood-Methode unter Verwendung des expectation maximization (EM) Algorithmus. Ein Voraussetzung ist eine Vorstellung über die Verteilung der betrachteten Variablen.

Alle bisher vorgestellten Tests sind nicht anwendbar auf die Situation, in der ausschließlich die Informationen für einen biallelischen Marker von einem erkrankten Kind und einem Elternteil vorhanden sind. Genau für diese Situation, in der die Verwendung des klassischen TDT gemäß CURTIS & SHAM (1995) nicht erlaubt ist, wurde von SUN *et al.* (1999) ein einfacher Ansatz vorgeschlagen, der die Verwendung dieser Informationen ermöglicht, ohne daß daraus ein Bias entsteht. Es wurden zwei Teststatistiken für Eltern-Kind Paare vorgestellt, die sowohl in familienbasierten Assoziationsstudien als auch in Kopplungsstudien gültige Test auf Kopplung und Assoziation sind. Die erste Statistik T_1 ist anwendbar, wenn eine der beiden folgenden Annahmen erfüllt sind. Die erste Annahme fordert Hardy-Weinberg Gleichgewicht, d.h. väterlichen und mütterlichen Genotypen sind am Markerlocus gleich verteilt. In der zweite Annahme fehlen Vater oder Mutter mit der gleichen Wahrscheinlichkeit von ½.

Die Statistik T_2 ist ein gültiger Test, wenn beide Annahmen verletzt sind. Der 1-TDT kann mit dem klassischen TDT zusammen berechnet werden. SUN *et al.* (1999) haben theoretisch und in Simulationen gezeigt, daß der 1-TDT das nominelle Signifikanzniveau von 5% hält und zu einem Powergewinn führt.

In dieser Arbeit wurde ein neuer Ansatz verfolgt, den fehlenden Genotypen eines Elternteiles mit Hilfe des expectation maximization (EM) Algorithmus (LAIRD, 1993) zu rekonstruieren. Die zentrale Idee dieses Verfahrens ist die Rekonstruktion, der fehlenden Genotypen unter Verwendung der vorhandenen Daten. Die Datenrekonstruktion wird mit dem EM-Algorithmus durchgeführt. Dazu wird in Schritt 1 zuerst eine Schätzung der Allel- und Genotypfrequenzen aus Eltern der vollständigen Familien vorgenommen. In Schritt 2 werden unter Verwendung dieser Frequenzen die Genotypen der fehlenden Eltern auf zwei verschiedene Weisen geschätzt. Bei der ersten Methode werden die beiden fehlenden Allele einzeln geschätzt, und bei der zweiten Methode werden die fehlenden Genotypen direkt geschätzt. Danach erfolgt in Schritt 3 eine Neuschätzung der Allel- und Genotypfrequenzen unter Verwendung aller elterlichen Daten, einschließlich der rekonstruierten Daten. Schritt 2 und 3 werden so oft wiederholt bis die Allel- bzw. Genotypfrequenzen hinreichend stabil geschätzt sind.
In einer ersten Variante werden die einzelnen Allele separat rekonstruiert (EM-Allelrekonstruktion) und in einer zweiten Variante den Genotyp direkt rekonstruiert (EM-Genotyprekonstruktion). Bei jedem Elter-Kind Paare wird danach geprüft, werde Genotypen für den fehlenden Elternteil möglich sind. Bei der EM-Allelrekonstruktion werden die Häufigkeiten der möglichen Genotypen unter Verwendung der geschätzten Allelfrequenzen berechnet, bei der EM-Genotyprekonstruktion werden die geschätzten Genotypfrequenzen zugrundegelegt. Der Eintrag in die T/NT-Tabelle erfolgt dann relativ zu den entsprechenden Genotypwahrscheinlichkeiten. Dieses Vorgehen ist möglich, da für den Eintrag in die T/NT-Tabelle und für die Berechnung der TDT-Statistik ganze Zahlen nicht zwingend notwendig sind. Die beiden neuen Varianten des EM-Rekonstruktions TDT wurden in Monte-Carlo Simulationen untersucht und mit dem 1-TDT verglichen.
Für die Monte-Carlo Simulationen wurde ein entsprechendes frei verfügbares Computerprogramm in der Computersprache GAUSS© geschrieben, mit dem die Generierung beliebig vieler Familientrios mit einem erkrankten Kind möglich ist. Der Zufallsgenerator verwendet die Systemzeit als Seed. Die notwendigen Familientrios

werden über die Wahrscheinlichkeiten aus Tabelle 3 einer Arbeit von KNAPP *et al.* (1993) unter der Vorgabe verschiedener Populationsparameter erzeugt. Bei einem beliebigen Anteil dieser Familien können die Genotypinformationen des Vaters gelöscht werden und mit Hilfe der beiden EM-Rekonstruktionsmethoden rekonstruiert werden. Zusätzlich wurde der 1-TDT in das Programm implementiert.

Die Monte-Carlo Simulationen wurden bei einer Gruppe von 100 Familien mit 10.000 Replikationen durchgeführt und basierten auf folgenden Annahmen: Die kompletten Genotypinformationen eines Elternteils fehlen zufällig. Es liegt Schichtung aus zwei Populationen vor. Der Anteil von Population 1 an der Gesamtpopulation variiert von 0 bis 100%. Beide Populationen sind im Hardy-Weinberg Gleichgewicht. Bei Assoziation zwischen Krankheit und Marker ist das Markerallel 1 mit der $P(1) = m$ mit dem Krankheitsallel D mit $P(D) = p$ positiv assoziiert. Für beide Populationen werden verschiedene Konstellationen der Allelfrequenzen untersucht. Als Vererbungsmodelle werden ein rezessives Modell, ein dominantes Modell, ein additives Modell ohne Phänokopie und ein dominantes Modell mit reduzierter Penetranz und Phänokopie betrachtet. Für Simulationen unter der Alternativhypothese H_1, d.h. Vorliegen eines Kopplungsungleichgewichtes, ist der Assoziationsparameter für Population 1 $\delta_1 = \delta_{max} = \min(p,m)$ und die Rekombinationsrate $\theta = 0$ oder $\theta = 0{,}01$. Population 2 führt zu Populationsstratifikation einen Assoziationsparameter von $\delta_2 = 0$. Die Nullhypothese H_0 wird für die Varianten a) $\delta_1 \neq 0$ und $\theta = 0{,}5$ b) $\delta_1 = \delta_2 = 0$ und $\theta = 0{,}5$ c) $\delta_1 = \delta_2 = 0$ und $\theta < 0{,}5$ simuliert. Zum Schluß werden verschiedene Anteile väterlicher Genotypen aus dem Datensatz gelöscht und mit dem EM-Algorithmus rekonstruiert. Power und Fehler 1. Art der EM-Genotyprekonstruktion und der EM-Allelrekonstruktion werden mit dem TDT für die kompletten Daten und dem TDT für die kompletten Fälle nach dem Löschen verglichen. Zusätzlich wird eine Vergleich mit dem 1-TDT unter den gleichen Bedingungen hergestellt.

In allen Simulationen halten der TDT für die kompletten Daten und der TDT für die kompletten Fälle erwartungsgemäß das nominelle Signifikanzniveau von 5%. Bei einem geringen Anteil fehlender Genotypen wird bei der EM-Allelrekonstruktion, der EM-Genotyprekonstruktion und beim 1-TDT ein empirisches Signifikanzniveau nahe 5% beobachtet. Bei einem größeren Anteil fehlender Genotypen zeigt die EM-Genotyprekonstruktion dagegen häufig ein deutlich erhöhtes empirisches Signifikanzniveau. Es ist also in einigen Fällen davon auszugehen, daß die EM-

Genotyprekonstruktion zu liberal ist. Sie kann daher keinesfalls für die Praxis empfohlen werden. Im Fall einer Anwendung müßte mit einem nicht zu tolerierenden Risiko gerechnet werden, ein falsch positives Ergebnis zu bekommen. Hingegen ist das empirische Signifikanzniveau der EM-Allelrekonstruktion bei einem hohen Anteil fehlender Daten deutlich geringer als das nominelle Signifikanzniveau von 5%. Aufgrund der Konservativität des Tests ist in dieser Situation davon auszugehen, daß Power bei der Rekonstruktion verschenkt wird. Im Vergleich mit der EM-Allelrekonstruktion, die häufig zu konservativ ist, und der EM-Genotyprekonstruktion, die oft zu liberalen Ergebnisse führt, schöpft der 1-TDT in allen MC-Simulationen das nominelle Signifikanzniveau gut aus. Beim Vergleich der Power zeigt sich, daß die EM-Genotyprekonstruktion deutlich besser abschneidet als die EM-Allelrekonstruktion oder der 1-TDT. Auf der Basis der Ergebnisse zum empirischen Signifikanzniveau ist diese Beobachtung nachvollziehbar. Da die EM-Genotyprekonstruktion im allgemeinen zu liberal ist, konnte ein höherer Powergewinn erwartet werden.

Bei der Beobachtung der empirischen Signifikanzniveaus für den 1-TDT liegen die Werte in allen simulierten Fällen im Bereich eines nominellen Signifikanzniveaus von 5%. Damit erfüllt der 1-TDT die optimalen Bedingungen für einen statistischen Test in Bezug auf das Risiko eines Fehlers 1. Art. Beim Vergleich der Powergewinn zwischen EM-Allelrekonstruktion und 1-TDT zeigt der 1-TDT regelmäßig etwas höhere Powerwerte. Da der 1-TDT in allen demonstrierten Fällen das nominelle Niveau von 5% halten kann, verschenkt er daher auch weniger Power als die EM-Allelrekonstruktion.

Für Simulationen mit zwei verschiedenen Markerallelfrequenzen in den beiden Subpopulationen verändert sich die Beobachtung zwischen den beiden EM-Rekonstruktionsmethoden nicht. Bei dem rezessiven Vererbungsmodell besteht die größte Power, eine bestehende Assoziation mit den Testmethoden nachzuweisen. Für die anderen Modelle nimmt die Power in der Reihenfolge, additiv dominante Vererbung, dominante Vererbung und dominante Vererbung mit reduzierter Penetranz und Phänokopie, ab. Die Beobachtungen zu den empirischen Signifikanzniveaus ist bei allen Vererbungsmodellen ähnlich. Die EM-Genotyprekonstruktion kann das Niveau nicht halten und ist zu liberal, während die EM-Allelrekonstruktion zu konservativ ist. Nur der 1-TDT hält in allen demonstrierten Fällen das Niveau.

Daher sollte bei Vorliegen von Hardy-Weinberg Gleichgewicht der 1-TDT gegenüber der EM-Allelrekonstruktion und der EM-Genotyprekonstruktion bevorzugt werden. Zum einen schöpft der 1-TDT das nominelle Signifikanzniveau von 5% aus und die Anwendung des Tests ist ohne aufwendige mathematische Berechnungen möglich.
Die Anwendung der Verfahren wurde in einer Reanalyse von zwei Kopplungs- und Assoziationsstudien zwischen Anorexia nervosa sowie Adipositas per magna und zwei Polymorphismen illustriert, wenn Daten von Trios und Paaren verwendet werden (HINNEY *et al.*, 1997a; HINNEY *et al.*, 1997b). Für beide genetischen Marker führte die Reanalyse zu identischen Ergebnissen wie die Originalpublikationen. Sowohl die EM-Rekonstruktionsverfahren als auch der 1-TDT konnten keine Assoziation und Kopplung zwischen dem Trp64Arg Polymorphismus des β_3-adrenergen Rezeptors bzw. dem 5-HTTLPR Polymorphismus in der Serontonin-Transport Region zur Regulation des Körpergewichtes nachweisen.

Für die Zukunft wäre interessant den 1-TDT auch auf multiallelische Marker zu erweitern. Dann könnten in einem Powervergleich Unterschiede zwischen dem 1-TDT und dem RC-TDT untersucht werden. Der Vorteil des 1-TDT liegt in seiner Eigenschaft, Genotypinformationen von kompletten Familien, inkompletten Familien mit Geschwisterkindern und inkompletten Familien ohne Geschwisterkinder zu kombinieren. Neue Ansätze für Fall-Kontroll Studien auch bei möglicher Populationsstratifikation wurden erst kürzlich von DEVLIN & ROEDER (1999) und PRITCHARD & ROSENBERG (1999) vorgestellt. Mit Hilfe von SNP-Chips („single nucleotide polymorphism“) kann sowohl auf Assoziation als auch auf Schichtung unter den Testpersonen gefahndet werden. SNP-Marker unterscheiden sich in ihrer DNA-Variante nur in einem einzigen Nucleotid. Dadurch ergibt sich der Vorteil einer automatisierten Analyse innerhalb einer kurzen Zeit. Allerdings besteht die Notwendigkeit, die Probleme durch multiples Testen und durch mögliche Fehltypisierung in Zukunft zu minimieren.

10 Zusammenfassung

Bei Assoziationsstudien zur Identifikation von Genen, die einen kleinen Beitrag zur Ausprägung einer Krankheit liefern, stellt ethnische Heterogenität der Studienpopulation (Populationsstratifikation) als confundierende Variable ein großes Problem dar. Der Transmission-Disequilibrium Test (TDT) (SPIELMAN *et al.*, 1993) vermeidet dieses Problem und ist bei Populationsstratifikation ein gültiger Test auf Assoziation. Das klassische Design des TDT ist einfach, da nur Markerallele innerhalb einer Kernfamilie mit einem erkrankten Kind und beiden Eltern untersucht werden. Der TDT testet, ob Eltern, die an einem biallelischen Markerlocus heterozygot sind, ein Allel bevorzugt an ihr erkranktes Kind weitervererben. Für diesen Test sind in seiner klassischen Form vollständig am Markerlocus typisierte Familientrios notwendig. Es resultiert ein Bias, wenn einzelne Elternteile am Markerlocus nicht typisiert werden können (CURTIS & SHAM, 1995). Zur Umgehung des Problems wurde der Sib-TDT (SPIELMAN & EWENS, 1998), der SDT (HORVATH & LAIRD, 1998) und der RC-TDT (KNAPP, 1999) vorgeschlagen, die beide Genotypinformationen von Geschwisterkindern benötigen. Von SUN *et al.* (1999) wurden zwei Teststatistiken, für Eltern-Kind Paare vorgestellt, die als 1-TDT bezeichnet werden, und auf einem Schätzer des Relativen Risikos basieren. Die erste Statistik kann angewandt werden, wenn die Genotypen bei beiden Eltern gleichermaßen verteilt sind, oder Vater und Mutter mit der gleichen Wahrscheinlichkeit fehlen.

In meiner Arbeit schlage ich ein neues Verfahren vor und rekonstruiere die Genotypen der fehlenden Elternteile mit Hilfe des expectation maximization (EM) Algorithmus (LAIRD, 1993). Die Anwendung des neuen EM-Rekonstruktions TDT wird durch Monte-Carlo Simulationen untersucht. Die Anwendung des Verfahrens wird für zwei Kopplungs- und Assoziationsstudie zwischen Anorexia nervosa sowie Adipositas per magna und zwei Polymorphismen illustriert, wenn Daten von Trios und Paaren verwendet werden (HINNEY *et al.*, 1997a; HINNEY *et al.*, 1997b).

Monte-Carlo Simulationen werden unter Verwendung der Computersprache GAUSS© durchgeführt. Die notwendigen Familientrios werden unter der Vorgabe verschiedener Populationsparameter mit den von KNAPP *et al.* (1993, Tabelle 3) beschriebenen Wahrscheinlichkeiten erzeugt. Es werden je 100 Familien mit 10.000 Replikationen generiert. Die Monte-Carlo Simulationen basieren auf folgenden Annahmen: Die komplette Genotypinformation eines Elternteils fehlt zufällig. Es liegt Schichtung aus

zwei Populationen vor. Der Anteil von Population 1 variiert von 0 bis 100%. Für beide Populationen gilt das Hardy-Weinberg Gleichgewicht. Bei Assoziation zwischen Krankheit und Marker sei das Markerallel 1 mit der Frequenz m positiv assoziiert mit dem Krankheitsallel D mit der Frequenz p. Für beide Populationen werden verschiedene Konstellationen der Allelfrequenzen untersucht. Als Vererbungsmodelle werden ein rezessives Modell, ein dominantes Modell, ein additives Modell ohne Phänokopie und ein additives Modell mit reduzierter Penetranz und Phänokopie berücksichtigt. Für Monte-Carlo Simulationen unter der Alternativhypothese H_1, d.h. Vorliegen eines Kopplungsungleichgewichtes, ist der Assoziationsparameter für Population 1 $\delta_1=\delta_{max}=\min(p,m)$ und die Rekombinationsrate $\theta=0$ oder $\theta=0{,}01$. Population 2 führt zu Populationsstratifikation bei einem Assoziationsparameter von $\delta_2=0$. Die Nullhypothese H_0 wird für die möglichen Varianten a) $\delta_1\neq 0$ und $\theta=0{,}5$ b) $\delta_1=\delta_2=0$ und $\theta=0{,}5$ c) $\delta_1=\delta_2=0$ und $\theta<0{,}5$ simuliert. Zum Schluß werden verschiedene Anteile väterlicher Genotypen aus dem Datensatz gelöscht und mit dem EM-Algorithmus rekonstruiert. Power und Fehler 1. Art der EM-Genotyprekonstruktion und der EM-Allelrekonstruktion werden mit dem TDT für die kompletten Daten und dem TDT für die kompletten Fälle nach dem Löschen verglichen. Zusätzlich werden die EM-TDTs dem 1-TDT unter den gleichen Modellen verglichen.

In allen Monte-Carlo Simulationen halten der TDT für die kompletten Daten und der TDT für die kompletten Fälle erwartungsgemäß das nominelle Signifikanzniveau von 5%. Die EM-Genotyprekonstruktion zeigt dagegen in einigen Fällen ein deutlich erhöhtes empirisches Signifikanzniveau, während bei der EM-Allelrekonstruktion tendenziell ein empirisches Signifikanzniveau von unter 5% gefunden wird. Der Fehler 1. Art liegt für den 1-TDT in allen Fällen nahe 5%. Bei der Powerberechnung werden für die EM-Genotyprekonstruktion die höchsten Werte gefunden. Die EM-Allelrekonstruktion und der 1-TDT zeigen beide eine nur wenig geringere Power als die EM-Genotyprekonstruktion. Dabei hat der 1-TDT im allgemeinen etwas höhere Power als der EM-Allelrekonstruktions TDT.

Die Statistik der EM-Genotyprekonstruktion ist insgesamt zu liberal, während die EM-Allelrekonstruktion zu konservativ ist, so daß dabei Power verschenkt wird. Aufgrund der gesicherten statistischen Eigenschaften und seiner größeren Power als der EM-TDT sollte dem 1-TDT bei Vorhandensein von Eltern-Kind-Paaren der Vorzug vor den anderen in dieser Arbeit diskutierten Testverfahren gegeben werden.

11 Literaturverzeichnis

Baron, M. (1997): Association studies in psychiatry: a season of discontent. *Mol Psychiatry* **2**: 278-281.

Blackwelder, W.C. & Elston, R.C. (1985): A comparison of sib-pair linkage tests for disease susceptibility loci. *Genet Epidemiol* **2**: 85-97.

Böddeker, I. & Ziegler, A. (2000): Assoziations- und Kopplungsstudien zur Analyse von Kandidatengenen. *Dtsch med Wschr* **125**: 810-815.

Christensen, R. (1996): *Analysis of variance, design and regression – Applied statistical methods*. London: Chapman & Hall.

Collins, F.S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R. & Walters, L. (1998): New goals for U.S. Human Genom Project: 1998-2003. *Science* **282**: 682-689.

Curtis, D. (1997): Use of siblings as controls in case-control association studies. *Ann Hum Genet* **61**: 319-333.

Curtis, D. & Sham, P.C. (1995): A note on the application of the transmission disequilibrium test when a parent is missing. *Am J Hum Genet* **56**: 811-812.

Devlin, B. & Roeder, K. (1999): Genomic control for associaton studies. *Biometrics* **55**: 997-1004.

Editorial (1999): Freely associating. *Nature Genet* **22**: 1-2.

Elston, R.C. (1998): Linkage and association. *Genet Epidemiol* **15**: 565-576.

Enache, D. (1994): *Numerische und statistische Verfahren zur Implementation künstlicher neuronaler Netze.* Diplomarbeit, Fachbereich Wirtschaftswissenschaften. Bergische Universität - Gesamthochschule Wuppertal.

Ewing, B. & Green, P. (2000): Analysis of expressed sequence tag indicates 35,000 human genes. *Nature Genet* **25**: 232-234.

Ewens, W.J. & Spielman, R.S. (1995): The transmission/disequilibrium test: history, subdivision, and admixture. *Am J Hum Genet* **57**: 455-464.

Falk, C.T. & Rubinstein, P. (1987): Haplotype relative risks: an easy way to construct a proper sample for risk calculations. *Ann Hum Genet* **51**: 227-233.

Ford, D., Easton, D.F., Stratton, M., Narod, S., Goldgar, D., Devilee, P., Bishop, D.T., Weber, B., Lenoir, G., Chang-Claude, J., Sobol, H., Teare, M.D., Struewing, J., Arason, A., Scherneck, S., Peto, J., Rebbeck, T.R., Tonin, P., Neuhausen, S.,

Barkardottir, R., Eyfjord, J., Lynch, H., Ponder, B.A., Gayther, SA., Zelada-Hedman, M., *et al.* (1998): Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium. *Am J Hum Genet* **62**: 676-689.

GAUSS© for Windows NT/95 (1997): Version 3.2.28. Maple Valley, WA.: Aptech Systems, Inc.

Hanis, C.L., Boerwinkle, E., Chakraborty, R., Ellsworth, D. L., Concannon, P., Stirling, B., Morrison, V.A., Wapelhorst, B., Spielman, R.S., Gogolin-Ewens, K.J., Shephard, J.M., Williams, S.R., et al. (1996): A genome-wide search for human non-insulin-dependent (type 2) diabetes genes reveals a major susceptibility locus on chromosome 2. *Nature Genet* **13**: 161-166.

Hinney, A., Barth, N., Ziegler, A., v. Prittwitz, S., Hamann, A., et al. (1997a): Serotonin transporter gene-linked polymorphic region: allele distributions in relationship to body weight and in anorexia nervosa. *Life Sciences* **61**: 295-303.

Hinney, A., Lentes, K.U., Rosenkranz, K., Barth, N., Roth, H., Ziegler, A., et al. (1997b): β3-adrenergic-receptor allele distributions in children, adolescents and young adults with obesity, underweight or anorexia nervosa. *Int J Obes* **21**: 224-230.

Horikawa, Y., Oda, N., Cox, N.J., Li, X., Orho-Melander, M., Hara, M., Hinokio, Y., Lindner, T.H., Mashima, H., Schwarz, P.E.H., del Bosque-Plata, L., Horikawa, Y., et al. (2000) : Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nature Genet* **26**: 163-175.

Horvath, S.M. & Laird, N.M. (1998): A discordant-sibship test for disequilibrium and linkage: no need for parental data. *Am J Hum Genet* **63**: 1886–1897.

Kaplan, N.L., Martin, E.R. & Weir, B.S. (1997): Power studies for the transmission/disequilibrium tests with multiple alleles. *Am J Hum Genet* **60**: 691-702.

Kennedy, W.J. & Gentle, J.E. (1980): *Statistical Computing*. Marcel Dekker: New York. S. 136-147.

Knapp, M., Seuchter, S.A. & Baur, M.P. (1993): The haplotype-relative-risk (HRR) method for analysis of association in nuclear families. *Am J Hum Genet* **52**: 1085-1093.

Knapp, M. (1999): The transmission/disequilibrium test and parental-genotype reconstruction: The reconstruction-combined transmission/disequilibrium test. *Am J Hum Genet* **64**: 861-870.

Kreienbrock, L. & Schach, S. (2000): *Epidemiologische Methoden.* 3. Auflage. Stuttgart: Gustav Fischer.

Laird, N.J. (1993): *Handbook of statistics*, Vol. 9. Amsterdam: Elvesier: S. 509-520.

Lander, E.S. & Kruglyak, L. (1995): Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genet* **11**: 241-247.

Lander, E.S. & Schork, N.J. (1994): Genetic dissection of complex traits. *Science* **265**: 2037-2048.

Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S.L. & Quackenbush, J. (2000): Gene index analysis of the human genome estimates approximately 120,000 genes. *Nature Genet* **25**: 239-240.

McGee, T.L., Devoto, M., Ott, J., Berson, E.L. & Dryja, T.P. (1998): Evidence that the penetrance of mutations at the RP11 locus causing dominant retinitis pigmentosa is influenced by a gene linked to the homologous RP11 allele. *Am J Hum Genet* **61**: 1059-1066.

Martin, R.B., Alda, M. & MacLean, C.J. (1998): Parental genotype reconstruction: applications of haplotype relative risk to incomplete parental data. *Genet Epidemiol.* **15**: 471-490.

Martin, R.B., Kaplan, N.L. & Weir, B.S. (1997): Tests for linkage and association in nuclear families. *Am J Hum Genet* **61**: 439-448.

Martin, E.R., Monks, S.A., Warren, L.L. & Kaplan, N.L. (2000): A test for linkage and association in general pedigrees: The pedigree disequilibrium test. *Am J Hum Genet* **67**: 146-154.

Martin. E.R., Bass, M.P. & Kaplan, N.L. (2001): Correcting for a potential bias in the pedigree disequilibrium test. *Am J Hum Genet* **68**: 1065-1067.

Morton, N.E. & Collins, A. (1998): Tests and estimates of allelic association in complex inheritance. *Proc Natl Acad Sci USA* **95**: 11389-11393.

Neffe, J. (1999): Die gentechnische Revolution. *Der Spiegel* **2**: 103-113.

Nimgaonkar, V.L. (1997): In defense of genetic association studies. *Mol Psychiatry* **2**: 275-277.

Nöthen, M.M., Propping, P. & Fimmers, R. (1992): Association versus linkage studies in psychosis genetics. *J Med Genet* **30**: 634-637.

Online Mendelian Inheritance in Man, OMIM (TM). John Hopkins University, Baltimore, MD. MIM Number {125853}:{21.12.2000}:. World Wide Web URL: http://www.ncbi.nlm.nih.gov/omim/.

Online Mendelian Inheritance in Man, OMIM (TM). John Hopkins University, Baltimore, MD. MIM Number {219700}:{21.9.2000}:. World Wide Web URL: http://www.ncbi.nlm.nih.gov/omim/.

Online Mendelian Inheritance in Man, OMIM (TM). John Hopkins University, Baltimore, MD. MIM Number {235200}:{4.12.2000}:. World Wide Web URL: http://www.ncbi.nlm.nih.gov/omim/.

Online Mendelian Inheritance in Man, OMIM (TM). John Hopkins University, Baltimore, MD. MIM Number {306700}:{17.1.2001}:. World Wide Web URL: http://www.ncbi.nlm.nih.gov/omim/.

Online Mendelian Inheritance in Man, OMIM (TM). John Hopkins University, Baltimore, MD. MIM Number {306900}:{10.1.2001}:. World Wide Web URL: http://www.ncbi.nlm.nih.gov/omim/.

Ott, J. (1989): Statistical properties of the haplotype relative risk. *Genet Epidemiol* **6**: 127-130.

Ott, J. (1991): *Analysis of human genetic linkage.* Revised edition. Baltimore: John Hopkins University Press.

Ott, J. (1996): Complex traits on the map. *Nature* **379**: 772-773.

Owen, M.J., Holmans, P. & McGuffin, P. (1997): Association studies in psychiatric genetics. *Mol Psychiatry* **2**: 270-273.

Paterson, A.D. (1997): Case-control association studies in complex traits – end of an era? *Mol Psychiatry* **2**: 277-278.

Pritchard, J.K. & Rosenberg N.A. (1999): Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet* **65**: 220-228.

Riede, U.N. & Schäfer, H.E. (1999): *Allgemeine und spezielle Pathologie*. 5. Auflage, Stuttgart; New York: Thieme.

Risch, N (1990): Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. *Am J Hum Genet* **46**: 242-253.

Risch, N. & Merikangas, K. (1996): The future of genetic studies of complex human diseases. *Science* **255**: 1516-1517.

Sachidanandam, R., Weissmann, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G. et al. (2001): A map of human sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**: 928-933.

Sachs, L. (1974): *Angewandte Statistik - Planung und Auswertung, Methoden und Modelle*. 4. Auflage, Heidelberg: Springer.

Schaid, D.J. & Sommer, S.S. (1994): Comparison of statistics for candidate-gene association studies using cases and parents. *Am J Hum Genet* **55**: 402-409.

Schepers, A. (1991): *Numerische Verfahren und Implementation der Schätzung von Mittelwert- und Kovarianzstrukturmodellen mit nichtmetrischen Variablen.* Dissertation, Wirtschaftswissenschaften. BUGH Wuppertal. Ahaus: Hartmann.

Siegenthaler, W. (1994): *Klinische Pathophysiologie*, 7. Auflage. Stuttgart: Thieme.

Spielman, R.S. & Ewens, W.J. (1996): The TDT and other family-based tests for linkage disequilibrium and association. *Am J Hum Genet* **59**: 983-989.

Spielman, R.S. & Ewens, W.J. (1998): A sibship test for linkage in the presence of association: sib transmission/disequilibrium test. *Am J Hum Genet **62: 450-458.***

Spielman, R.S., McGinnis, R.E. & Ewens, W.J. (1993): Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* **52**: 506-516.

Strachan, T. & Read, A.P. (1999): *Human molecular Genetics.* 2nd Edition, New York: Wiley-Liss.

Strickberger, M.W. (1988): *Genetik*. München: Hanser.

Sun, F., Flanders, W.D., Yang, Q. & Khoury, M.J. (1998): A new method for estimating the risk ratio. *Am J Epidemiol* **148**: 902-909.

Sun, F., Flanders, W.D., Yang, Q. & Khoury, M.J. (1999): Transmission disequilibrium test (TDT) when only one parent is available: The 1-TDT. *Am J Epidemiol* **150**: 97-104.

Wang D.G., Fan, J.B., Siao, C.J., et al. (1998): Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**: 1077-1082.

Weeks, D.E. & Lathrop, G.M. (1995): Polygenic disease: methods for mapping complex disease traits. *TIG* **11**: 513-519.

Weinberg, C.S. (1999): Allowing for missing parents in genetic studies of case-parent triads. *Am J Hum Genet* **64**: 1186-1193.

White, R., Lalouel, J.M. (1997): Kartierung von Chromosomen mit DNS-Markern. *Spektrum der Wissenschaft. Digest: Gene und Genome* **6**: 8-17.

Woolf, B. (1955): On estimating the relationship between blood group and disease. *Ann. Hum. Genet* **19**: 251-253.

Ziegler, A. & Hebebrand, J. (1998): Sample size calculations for linkage analysis using extreme sib pairs based on segregation analysis with the quantitative phenotype body weight as example. *Genet Epidemiol* **15**: 577-593.

12 Anhang

12.1 MC-Simulationsergebnisse und MC-Simulationsprogramm

Die vollständigen Ergebnisse der Monte-Carlo Simulationen und das Monte-Carlo Simulationsprogramm sind auf folgender Internetseite einsehbar:

URL: http//:www.j-mittemeyer.de

12.2 Verzeichnis meiner akademischen Lehrer

Mein akademischer Lehrer waren die nachfolgenden Damen und Herren

In Marburg

Arnold	Grzeschik	Lang	Schneider
Aumüller	Habermehl	Lange	Schulz
Basler	Happle	Lennartz	Seifart
Baum	Hasilik	Lorenz	Seitz
Bertalanffy	Hebebrand	Maisch	Seyberth
Bien	Hoffmann	Moosdorf	Slenczka
Daut	Joseph	Mutters	Steininger
Engel	Kalbfleisch	Oertel	Sturm
Fruhstorfer	Kälble	Petermann	Vohland
Fuhrmann	Kern	Rehder	Voigt
Ganz	Kleine	Remschmidt	Weihe
Gemsa	Klenk	Richter	Werner
Geus	Klose	Röhm	Wesemann
Gotzen	Kretschmer	Rothmund	Westermann
Gressner	Krieg	Schäfer	Wichert
Griss	Koolman	Schachtschabel	Ziegler

In Kassel

Faß	Kuhn	Raible	Sons
Fischer	Neuhaus	Schmidt	Tönnis
Hirschmann	Pausch	Schürmann	Vogt

In Luzern

Berberat	De Simoni	Staubli

12.3 Danksagung

Mein Dank gilt all denjenigen, die es mir ermöglicht haben, meine Dissertation an der Philipps-Universität Marburg erfolgreich anzufertigen.

Mein besonderer Dank gilt vor allem PD Dr. Andreas Ziegler für die Überlassung des Themas. Er hat mich bei der Durchführung dieser Arbeit kontinuierlich unterstützt und war jederzeit für Fragen offen. Ebenso danke ich Prof. Dr. Helmut Schäfer dem Direktor des Institutes für Medizinische Biometrie und Epidemiologie der Philipps-Universität Marburg.

Ich danke meinen Eltern Barbara und Rolf, die mich in jeder Hinsicht in meiner akademischen Ausbildung gefördert haben.

Meinem Onkel Dr. Peter Hocks möchte ich für das Korrekturlesen meines Manuskriptes danken.